# Automatic domain module extraction using SVM machine learning approach from electronic textbooks

**C. Vishnu priya[1], S.V.Hemalatha[2]**

PG Scholar, Department of CSE, Kalaignar Karunanidhi Institute of Technology,
Coimbatore, Tamilnadu, India[1]

Assistant Professor, Department of CSE, Kalaignar Karunanidhi Institute of Technology,
Coimbatore, Tamilnadu, India[2]

**Abstract-** In recent days, Technology-Supported Learning Systems (TSLSs), such as Intelligent Tutoring Systems (ITSs), Adaptive Hypermedia Systems (AHSs), and especially, Learning Management Systems (LMSs) are being extensively used in many studious institutions and becoming necessary for learning. The Domain Module is measured the core of any TSLSs as it represents the data about a topic matter to be communicated to the learner. In the existing system, a DOM-Sortze is a system that uses NLP techniques, heuristic analysis, and ontologies for the semiautomatic structure of the Domain Module from electronic textbooks. But in this system, still lack in the identification of pedagogical relationships. This is needed to improve in this system. In other words, DOM-Sortze system is not able to including the new rules of the pedagogical relationships. To overcome this issue, using learning techniques to learn the new rules in the pedagogical relationships. In our proposed system, we are proposing the SVM (support vector machine) learning approach intended for learning process. Our machine learning methods are used to infer new rules in order to improve the identification of pedagogical relationships or the DRs in the electronic textbooks.

**Keywords:** knowledge acquisition, SVM, domain extract, ontology learning

## I. INTRODUCTION

These days, large quantity of data is being accumulated in the data depot. Usually there is a huge gap from the stored data to the information that could be constructed from the data. This changeover won't occur routinely, that's where Data Mining comes into image. In investigative Data Analysis, some initial knowledge is known about the data, but Data Mining could help in a more in-depth knowledge about the data. in search of knowledge from enormous data is one of the most required attributes of Data Mining. Labour-intensive data analysis has been approximately for some time now, but it creates a blockage for huge data analysis.

Data stored in text databases is mainly semi-structured, it is neither totally formless nor totally prepared. Without expressive what could be in the credentials, it is difficult to invent successful queries for analyzing and extracting useful information from the data. Users need tools to compare different documents, rank the importance and relevance of the documents, or discover patterns and trends across various documents. Thus, Text Mining has become an ever more accepted and indispensable theme in Data Mining.

Data are any facts, numbers, or text that can be processed by a computer. nowadays, organizations are accumulating infinite and growing amounts of data in dissimilar formats and dissimilar databases. Information can be transformed into knowledge about historical patterns and future trends.

Knowledge discovery in databases process, or KDD is comparatively young and interdisciplinary pasture of computer science is the process of discovering new patterns from large data sets concerning methods at the connection of artificial intelligence, machine learning, statistics and database systems. The goal of data mining is to extract knowledge from a data set in a human-understandable structure. Datamining is the entire process of applying computer-based methodology, including new techniques for knowledge discovery, from data. Databases, Text Documents, Computer Simulations, and Social Networks are the Sources of Data for Mining.

Building the Domain Module is a hard task which entails not only selecting the domain topics to be learned, but also defining the pedagogical relationships among the topics that determine how to plan the learning sessions. Textbook authors deal with similar problems while writing their documents, which are structured to facilitate comprehension and learning. Artificial intelligence techniques provide the means for the semiautomatic construction of the Domain Modules from electronic textbooks which may significantly contribute to reduce the development cost of the Domain Modules. DOM-Sortze

DOI 10.17148/IJARCCE.2015.4464

aims to be domain independent, so no domain-specific knowledge is used except the processed electronic textbook and the knowledge gathered from it.

## II. TEXT MINING

The prepared input data lacked numeric or categorical fields suitable for traditional clustering and classification models. Instead, the text descriptions of pump station maintenance jobs were transformed into vectors of term weights, which could then be used for clustering and classification. The main motivation of our research is to study different existing tools and techniques of Text Mining for Information Retrieval (IR). Search engine is the most well known Information Retrieval tool. Application of Text Mining techniques to Information Retrieval can improve the precision of retrieval systems by filtering relevant documents for the given search query. Electronic information on Web is a useful resource for users to obtain a variety of information.

Given below are some issues identified in Information Retrieval process: Traditional Information Retrieval techniques become inadequate to handle large text databases containing high volume of text documents. To search relevant documents from the large document collection, a vocabulary is used which map each term given in the search query to the address of the corresponding inverted file; the inverted files are then read from the disk; and are merged, taking the intersection of the sets of documents for AND, OR, NOT operations. To support retrieval process, inverted file require several additional structures such as document frequency of each lexicon in the vocabulary, term frequency of each term in the document.

Text mining, also known as intellectual Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers usually to the process of extracting attractive and non-trivial information and knowledge from amorphous text. Data may be discovered from many resources of information. Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files.

## III. RELATED WORKS

Classification is one of the most important tasks for different application such as text categorization, tone recognition, image classification, micro-array gene expression, proteins structure predictions, data Classification etc. The Support Vector Machine (SVM) was first proposed by Vapnik and has since attracted a high degree of interest in the machine learning research community. In this work, a novel learning method, Support Vector Machine (SVM), is applied on different data (Diabetes data, Heart Data, Satellite Data and Shuttle data) which have two or multi class. SVM, a powerful machine method developed from statistical learning and has made significant achievement in some field. Introduced in the early 90's, they led to an explosion of interest in machine learning. The foundations of SVM have been developed by Vapnik and are gaining popularity in field of machine learning due to many attractive features and promising empirical performance. SVM method does not suffer the limitations of data dimensionality and limited samples[1].

The objective of business intelligence (BI) is to make well-informed business decisions by building both succinct and accurate models based on massive amounts of practical data. Support vector machines (SVM) have been applied to build classifiers, which can help users make well-informed business decisions. Despite their high generalisation accuracy, the response time of SVM classifiers is still a concern when applied into real-time business intelligence systems, such as stock market surveillance and network intrusion detection. This work speeds up the response of SVM classifiers by reducing the number of support vectors. Based on the above motivation, this paper proposes a new algorithm called K-means SVM (KMSVM). The KMSVM algorithm reduces support vectors by combining the K-means clustering technique and SVM. Since the K-means clustering technique can almost preserve the underlying structure and distribution of the original data, the testing accuracy of KMSVM classifiers can be under control to some degree even though reducing support vectors could incur a degradation of testing accuracy[2].

The classification of medical data has become an increasingly challenging problem, due to recent advances in medical mining technology. Classification of this tremendous amount of data is time consuming and utilizes excessive computational effort, which may not be appropriate for many applications. In this work, we develop an approach to optimize the support vector machine parameters which combines the merits of support vector machine (SVM) and genetic algorithm (GA). This work focuses on the combining a feature selection technique based on genetic algorithm and support vector machines (SVM) of medical disease classification. SVMs have been used for various applications as a powerful tool for pattern classification. We use evolutionary computation which is a subfield of artificial intelligence or computational intelligence that involves combinatorial optimization problems. Evolutionary computation uses iterative progress, such as growth or development in a population. This population is then selected in a guided random search using parallel processing to achieve the desired end. It have used the Weka toolkit to experiment with these five data mining algorithms. The Weka is an ensemble of tools for data classification, regression, clustering, association rules, and visualization. WEKA version 3.6.9 was utilized as a data mining tool to evaluate the performance and effectiveness of the SVM and Proposed SVM-Genetic technique[3].

In this work, a novel multilayered neuro-fuzzy classifier is suggested, the self-organizing neuro-fuzzy multilayered

DOI 10.17148/IJARCCE.2015.4464

classifier (SONeFMUC), incorporating the principles of classifiers combination, decision fusion and feature transformation. In regard to other classification methods of the literature, this approach provides the following innovations and distinctions:

(1) The SONeFMUC classifier introduces a new class of hierarchical classifiers comprising a number of elementary fuzzy neuron classifiers (FNCs) arranged in layers. At each layer, two parent FNCs are combined to produce a descendant FNC at the next layer with better classification capabilities. Currently, hierarchical classifiers are confined to combining large-scale classifiers arranged in a single layer. SONeFMUC extends the above idea by continuing combination of small-scale classifiers for multiple layers.

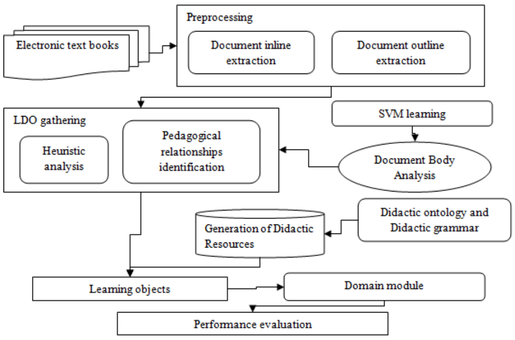
(2) A salient asset of SONeFMUC lies on the FNC structure. Generally, each FNC is composed of four modules: the fuser, the data splitting (DS) module, the fuzzy partial description (FPD) and the decision making fuzzy unit (DMFU). The fuser aggregates the decision outputs of the parent classifiers while DS divides data set into wellclassified patterns and ambiguous ones. The last two modules realize a neuro-fuzzy classifier within each FNC, used to improve the classification accuracy of ambiguous patterns. Accordingly, our model incorporates the notions of the voting/rejection scheme in a more effective way at the FNC level[4].

In this work, a new approach for data-driven incremental learning for Takagi–Sugeno fuzzy systems is demonstrated, called flexible fuzzy inference systems (FLEXFIS). One main focus lies on a reasonable connection between the adaptation of nonlinear premise parameters and linear consequent parameters. This is for the purpose of obtaining a solution for the linear consequent parameters as close as possible to the optimal one in the least-squares sense, i.e., when using the mean-squared error as an underlying optimization function (note: optimality can be analytically achieved when using batch training;). In this context, we call this a robust incremental/ evolving training procedure. For doing so, the idea of vector quantization, in combination with a so-called vigilance parameter motivated, is exploited to form a modified version of VQ, called VQ-INC-EXT. This novel clustering approach finds the nearest (winning) cluster by calculating the distance from a new data point to the surface (instead of to the centers, as in conventional VQ) of already obtained clusters. The surface is, sample wise, updated synchronously to the cluster centers. Furthermore, the new approach builds up clusters in an incremental manner without the need for pre parameterizing the number of clusters. By projecting clusters onto the input axes after each incremental learning step, it is applicable for online training of fuzzy partitions as well as rule bases[5].

## IV. PROBLEM STATEMENT

- Identification of pedagogical relationships has low accuracy rate.
- Don't have knowledge about the new rules or updated rules in the pedagogical relationships or didactic resources (DRs).
- DOM-Sortze system is not able to including the new rules of the pedagogical relationships. Thus, the overall performance of the system is degrades in this system.
- Objectives
- To improve the accuracy of the system, thus undoubtedly the performance of the system is improved.
- To enhance the effectiveness of the system compared to the existing system.
- To obtain the system that can able to learn the new rules or updated rules in the pedagogical relationships or didactic resources (DRs) accurately. Therefore, the accuracy of Identification of pedagogical relationships is increased.

## OVERALL DESIGN


Electronic text books
Preprocessing
Document inline extraction
Document outline extraction
LDO gathering
Heuristic analysis
Pedagogical relationships identification
SVM learning
Document Body Analysis
Didactic ontology and Didactic grammar
Generation of Didactic Resources
Learning objects
Domain module
Performance evaluation

## V. IMPLEMENTATION

### A. PREPROCESSING

The system prepares the electronic document and gathers a consistent depiction of it, to later run the knowledge attainment processes. The electronic documents are specified in many formats such as pdf(Portable Document Format), rtf(Rich Text Format), doc, or odf(Open Document Format). If the preprocess is carried out first to prepare the document. In that preprocessing has to convert the input document into Text format. After converting the document it can separate the input as document inline extraction and document outline extraction and to build the hierarchical structure for analysing the part-of-speech information that will be used in further steps for the domain extract.

### B. LDO GATHERING

The LDO contains the main domain topics and the pedagogical relationships among them. LDO means Learning Domain Ontology. In that pedagogical relationships can be divided into two ways as structural and sequential. If the structural relationship can be performed isA and partof relationship and the sequential relationship can be performed prerequisite and next relationship. The L isA D relationship declares the topic L is a particular kind of D. The L partof D defines that L is a partof D, then L is one of the topics to learn completely master D. The D prerequisite L relationship denotes that a

DOI 10.17148/IJARCCE.2015.4464

topic L must be mastered before perform to learn topic D. If L next D denotes that to learn topic D right after mastering topic L. The LDO describes a certain domain for learning purposes. In the LDO gathering process an internal representation is used. The LDO gathering entails two main NLP and heuristic analysis based steps: outline analysis and Document body analysis.

1) Outline analysis:
The outline analysis describes the results in an initial version of the Learning Domain Ontology. The outline analysis is derived of two phases such as Basic analysis and Heuristic analysis.
i) Basic analysis:
In this mission the main topics of the domain and the pedagogical relationships in the middle of these topics are mined from the homogenized outline internal representation. In the basic analysis each file item is measured as a domain topic. The construction of the manuscript outline is used as a means to collect pedagogical relationships. A subitem of a general topic is used to define part of it. as a result, structural relationships are defined between every outline item and all subitems. The order of the outline items reflects the suggested sequence for learning the domain topics. hence an original set of sequential relationships is acknowledged from the instruct of the outline items.

ii) Heuristic analysis:
The results of the basic analysis are sophisticated based on a set of heuristics that both classify the relationships acknowledged in the preceding step and also extract new ones. It involve the condition to be coordinated, the empirically gathered assurance on the heuristic and the postcondition. In the heuristic analysis contain two steps as the heuristics for identifying structural and sequential relationships. For that structural can be classified as Individual and Group structural heuristics. If the individual structural relationships contains different steps as MWH, ENH, AH, He-MWH, A+MWH, PGH and also the group structural relationship contains KH, CHe+MWH. After that the sequential relationships perform RH, He+RH, A+RH, PGH, He+PGH, A+PGH.

2) Document body analysis:
It enhances the ontology with new topics and relationships.
i) Identifying New Topics:
For identifying new topics enhancing the LDO gathered in the previous stage with new domain topics. It is used to identify only the new topics in a given document.

ii) Identifying New Relationships among Topics:
This process allows the recognition of new pedagogical relationships from the electronic document using pattern based approach. The grammar contains a set of rules recitation syntactic structures equivalent to pedagogical relationships.

### C. GENERATION OF THE DRs

The recognition of the DRs is approved out by identifying pertinent text remains that communicate to definitions, examples, theories, principles, and problem statements for the LDO topics. Given that the DRs recognized by the DR grammar are usually quite simple, they are enhanced in two ways to make them more accurate. On the one hand, consecutive DRs are combined if they are similar, to which end similarity measures have been defined. Two consecutive atomic definitions that might be combined to get a more comprehensive DR.

### D. SVM LEARNING

The machine learning methods which is used for effectively identify the pedagogical relationships. In this approach, we are learning the new rules in the pedagogical relationships or the DRs in the electronic textbooks in the training phase. Based on the training phase, we can easily identify the pedagogical relationships or the DRs in the electronic textbooks. Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships, the SVM modeling algorithm finds an optimal hyperplane with the maximal margin to separate two classes, which requires solving the following optimization problem.
Maximize

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j\, y_i y_j k(x_i, x_j)$$

Subject to,

$$\sum_{i=1}^{n} \alpha_i\, y_i = 0$$

Where $0 \le \alpha_i \le b \; i = 1,2, \ldots \ldots, n$ Where $\alpha_i$ is the weight of training sample $x_1$. If $\alpha_i >$ 0, x1 is called a support vector b is a regulation parameter used to trade-off the training accuracy and the model complexity so that a superior generalization capability can be achieved. K is a kernel function, which is used to measure the similarity between two samples. A popular radial basis function (RBF) kernel functions. This process is repeated k times for each subset to obtain the cross validation performance over the whole training dataset. If the training dataset is large, a small subset can be used for cross validation to decrease computing costs. The following algorithm can be used in the classification process.

Input: sample x to classify training set T = $\{(x_1,y_1),(x_2,y_2),\ldots\ldots(x_n,y_n)\}$; number of nearest neighbours k.

Output: decision $y_p$ Î {-1,1}
Find k sample $(x_i,y_i)$ with minimal values of $K(x_i,x_i) - 2 * K(x_i,x)$

Train an SVM model on the k selected samples Classify x using this model, get the result $y_p$ Return $y_p$

DOI 10.17148/IJARCCE.2015.4464

## ALGORITHM

Input: Number of the training samples (determined in existing system) with dataset w as input data point for

SVM classification
Output: Classification result i.e., prediction of the pedagogical relationships result
Procedure SVM (w)
Begin
Initialize C=0

Get input file dataset w for training
Read the number of input training dataset W from original dataset
$x_i.w + b = 0$
$x_i.w + b = 1$
Decision function $f(W) = x_i.w - b$
If $f(W) \geq 1$ for $x_i$ is the first class ; Else
$f(W) \leq -1$ for $x_i$ is the second class

The prediction result for (i=1,…n) number of training samples
$y_i(x_i.w - b) \geq 1$
Display the result

E. LOs GATHERING

The generation of LOs for the domain topics is achieved by identifying and gathering DRs. LO means Learning Objects. The LDO and OWL ontology are used to build the LOs from the gathered DRs. LOs are stored in the LOR (Learning Objects Repository) for further use.

F. PERFORMANCE EVALUATION

Precision: Precision value is calculated is based on the retrieval of information at true positive prediction, false positive. Data precision is calculated the percentage of positive results returned that are relevant.

Recall: Recall value is calculated is based on the retrieval of information at true positive prediction, false negative. Recall is the fraction of relevant instances that are retrieved.

F-measure: F-measure is the harmonic mean of precision and recall.

$$F = 2. \frac{\text{Precision. recall}}{\text{Precision + recall}}$$

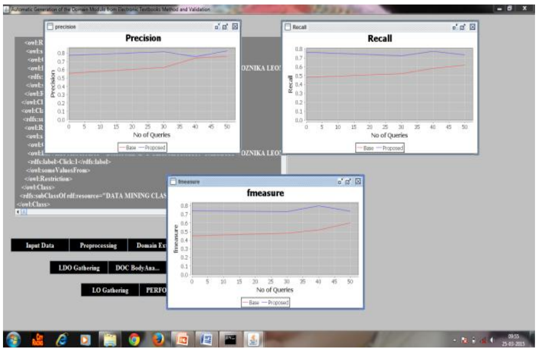


Fig: Performance evaluation
(Precision, Recall, Fmeasure)

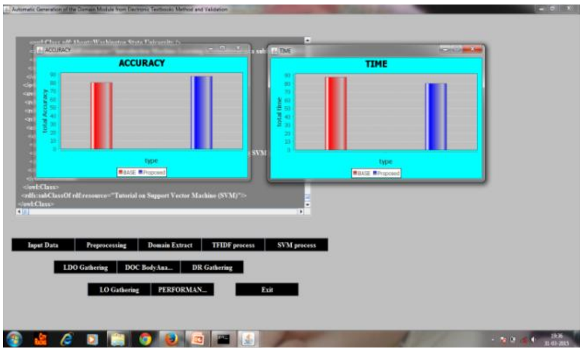


Fig: Accuracy and Time

## VI. CONCLUSION

In the previous work has presented DOM-Sortze, a system for the semiautomatic generation of the Domain Module from electronic textbooks. The system employs NLP techniques, heuristic reasoning, and ontologies for the knowledge acquisition processes. Our proposed system, introduce the machine learning methods as SVM which is used for effectively identify the pedagogical relationships. In this approach, we are learning the new rules in the pedagogical relationships or the DRs in the electronic textbooks in the training phase. Based on the training phase, we can easily identify the pedagogical relationships or the DRs in the electronic textbooks. Implementation results shows that our proposed system is well effective than the existing system. Our system is improves the accuracy of the system, thus undoubtedly the performance of the system is improved.

## REFERENCES

[1]. J.S. Justeson and S.M. Katz, "Technical Terminology: Some Linguistic Properties and an Algorithm for Identification of Terms in Text," Natural Language Eng., vol. 1, no. 1, pp. 9-27, 1995.

[2]. I. Alegria, A. Gurrutxaga, P. Lizaso, X. Saralegi, S. Ugartetxea, and R. Urizar, "An XML-Based Term Extraction Tool for Basque," Proc. Fifth Int'l Conf. Language Resources and Evaluations (LREC '04), 2004.

[3]. "Constraint Grammar: Language-Independent System for Parsing Unrestricted Text," Natural Language Processing, F.Karlsson, A. Voutilainen, and J. Heikkila, eds., no. 4, Mouton de Gruyter, 1995.

[4]. M. Larran˜ aga, I. Calvo, J.A. Elorriaga, A. Arruarte, K. Verbert, and E. Duval, "ErauzOnt: A Framework for Gathering Learning Objects from Electronic Documents," Proc. 11th IEEE Int'l Conf. Advanced Learning Technologies (ICALT '11), pp. 656-658, 2011.

[5]. T. Leidig, "L3-Towards an Open Learning Environment," ACM J. Educational Resources in Computing, vol. 1, no. 1, pp. 5-11, 2001.

[6]. K. Verbert, D. Ga_sevi_c, J. Jovanovi_c, and E. Duval, "Ontology-Based Learning Content Repurposing," Proc. 14th Int'l Conf. World Wide Web (WWW '05), pp. 1140-1141, 2005.

[7]. M. Larran˜ aga, A. Conde, I. Calvo, A. Arruarte, and J.A. Elorriaga, "Evaluating the Automatic Extraction of Learning Objects from Electronic Textbooks Using Erauzont," Proc. 11th Int'l Conf. Intelligent Tutoring Systems (ITS '12), pp. 655-656, 2012.

DOI 10.17148/IJARCCE.2015.4464