

## A Multilevel Semantic Document Classifier Based On SVM Integrated With Domain Ontologies

Vijayasundaram Uma<sup>\*1</sup>, Punnaivanam Sankar<sup>\*2</sup>, Gnanasekaran Aghila<sup>\*3</sup>

<sup>\*1</sup>*Department of Information Technology, Sri Manakula Vinayagar Engineering College  
Puducherry-605107, India*

<sup>\*2</sup>*Department of Chemistry, Pondicherry Engineering College Puducherry-605014, India*

<sup>\*3</sup>*Corresponding Author Department of Computer Science, Ramanujan School of Maths & CS  
Pondicherry University, Puducherry-605014, India*

*umabskr@gmail.com, gapspec@yahoo.com, aghilaa@yahoo.com*

### Abstract

*A multilevel semantic document classification system based on Support Vector Machine (SVM) in association with domain ontologies has been developed. The documents related to the scientific domains such as computer science and chemistry are treated as the test source. The classification results are more precise and fine grained when compared to the conventional methodologies. The sharpness of the classification has been found to be enhanced when the domain knowledge in terms of ontologies is integrated with SVM procedures. So the developed system provides the advantages of high generalization performance, prevention of over fitting, less computational complexity, high accuracy, and robustness. The use of automated identification of the semantic components derived from the domain ontologies enables the system to provide semantically rich classification results.*

### Keywords

*Document Classification, Support Vector Machine, Ontology, Semantic components, Text Mining.*

### 1. Introduction

The semantic document classification system is vital in many contexts such as document indexing based on a controlled vocabulary, document filtering, automated metadata generation, word sense disambiguation and population of hierarchical catalogues of web resources. In general for any application involving document organization or

selective and adaptive document dispatching, the semantic knowledge support at possible levels results in the efficient functioning of the system with meaningful outputs. This approach is particularly important in the classification of scientific documents which demands an increased level of granularity and semantic support for the precise results.

The conventional text categorization is a classification process applied to the textual domain to solve the problem of assigning text content to predefined categories. Due to the explosive availability of the information content both in public and corporate domains, the efficiency of organizing the text content with simple text categorization approach is not sufficient. Although SVM is the state-of-the-art classification algorithm enabling a powerful supervised learning paradigm [1], it cannot provide a semantic support during the scientific content classification. SVM classifiers create a maximum margin hyper plane that lies in a transformed input space and splits the example classes. Maximizing the distance to the nearest cleanly split examples results in a higher level precise classification [2]. However, it is not sufficient to handle the text contents composed with same term referring to different meanings and different terms giving the same meaning. This is because consideration of vocabulary with equivalent synonyms and generic terms is at most essential in order to capture all the relevant documents related to a particular search criterion.

The only way to face this situation is to integrate the SVM based procedures with the ontologies developed in the concerned domains. Generally, formal ontology deals with the interconnections of things, with objects and properties, parts and wholes, relations and collectives. As ontology has unique, hierarchical structure and provides support for machine reasoning starting from very primitive terms

[3, 4], the integration of the appropriate concept taxonomies with SVM is a novel methodology to develop classification systems with improved accuracy. Such ontology-driven classification is a powerful technique which combines the advantages of modern classification methods with semantic specificity of the ontologies.

In this paper the features of SVM are used for the classification of documents at multi-level. The classifications are subjected to the screening with the relevant ontologies to refine the categories into more precise and semantically rich classifications. This approach also results in a fast and accurate classification of documents. The documents are often organized around textual and graphical contents; the semantic components may or may not be explicitly indicated with basic structural elements [5]. In order to obtain a precise document classification the semantic components of the terms are also identified. A semantic component instance of the terms in any sections of the text document is captured by linking the filtered terms with the relevant concept taxonomies through SVM. Accordingly the terms containing the information about a particular aspect of an instance of a concept that is important in a domain can also be captured for the classification. For example in the chemistry domain, the term 'ethyl alcohol' has an equivalent term 'ethanol' and a formula 'C<sub>2</sub>H<sub>5</sub>OH' referring to the same meaning. The proposed methodology has the provision to check the equivalent requisites of the terms captured and the respective documents are included without being thrown before classification.

The collection of semantic components helps in identifying the respective categories to which the document belongs and avoids the document being classified under irrelevant categories. Information about semantic components in documents is identified for information retrieval in two phases. In the first phase, the documents are being subjected to classification with SVM using the domain knowledge, and are being classified in the appropriate location within a concept hierarchy belonging to a particular domain. In the second phase the semantic components existing in the resultant documents are collected along with the corresponding terms and the documents are reclassified precisely and semantically. Additionally, the list of the semantic components collected in each document provides a short synopsis of document content. This allows the possibility for a searcher to specify which semantic components are of interest, and even search for terms within semantic components.

The contents of this article are organized as follows. Section 2, briefly introduces the related work in text categorization. In Section 3, an introduction to SVM

has been provided. In Section 4, the system architecture and the major modules involved in this method for document classification has been presented. Section 5 describes the implementation details of document classification. In Section 6 the experimental results are discussed to evaluate the approach. Section 7 has the concluding remarks followed by the references.

## 2. Related Work

The automated Text categorization (or Document classification) into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the ensuing need to organize them.

A growing number of statistical learning methods have been applied to this problem in recent years, including regression models [6], nearest neighbor classifiers [7], Bayesian probabilistic classifiers [8,9], decision trees [6], inductive rule learning algorithms [10], neural networks [11], on-line learning approaches [12] and Support Vector Machines[2]. An Evaluation Of Statistical Approaches to Text Categorization has been performed[13].

However, generally machine learning methods over fit the training data when many features are given [14]. Hence techniques are required to concentrate in selection of features in order to overcome the deficiency mentioned. Support Vector Machines (SVMs) [2] are robust even when the number of features is large. Therefore, SVMs have shown good performance for text categorization [2], chunking [15].

In spite of these advantages SVM has the disadvantage of very high training time [16]. The training time for document classification can be considerably reduced by having background knowledge and hence a more comprehensive approach has been developed using the background knowledge available in the ontology. Ontology means information used in a specific domain and relationships defined in relation to the information. In this work domain ontology for computer science domain is constructed from Dmoz directory hierarchy. The process of document classification basically involves two procedures: Finding key vocabulary in the documents and mapping onto a node in the concept hierarchy (ontology) using the extracted words [4].

The advantages of an ontology-based classification approach over the existing ones, such as hierarchical and probabilistic approach [4], are. The nature of the relational structure of ontology provides a mechanism to enable machine reasoning.

- (1)The conceptual instances within ontology are not only a bag of keywords but have inherent

semantics and a close relationship with the class representatives of the classification schemes. Hence, they can be mapped to each other.

(2) It also enables getting insight into and observes the way the classifier assigns a class representative to a document by tracking the links between the conceptual instances involved and the associated class representative.

One useful application for automatic categorization is to support effective text retrieval. The automatically assigned categories will improve the retrieval performance compared to no categorization [17].

### 3. SVM

SVM is a supervised learning algorithm for 2-class problems. Training data is given by

$$(x_1, y_1), \dots, (x_u, y_u),$$

$$x_j \in \mathbb{R}^n, y_j \in \{+1, -1\} \quad (1)$$

Here,  $x_j$  is a feature vector of the  $j$  th sample;  $y_j$  is its class label, positive (+1) or negative (-1). SVM separates positive and negative examples by a hyper plane defined by

$$w \cdot x + b = 0, w \in \mathbb{R}^n, b \in \mathbb{R} \quad (2)$$

Figure 1 shows a linearly separable case. The SVM determines the optimal hyper plane by maximizing the margin. A margin is the distance between negative examples and positive examples. Since training data is not necessarily linearly separable, slack variables ( $\xi_j$ ) are introduced for all  $x_j$ . These  $\xi_j$  incur misclassification error, and should satisfy the following inequalities:

$$\begin{aligned} w \cdot x_j + b &\geq 1 - \xi_j \\ w \cdot x_j + b &\leq -1 + \xi_j \end{aligned} \quad (3)$$

Under these constraints, the following objective function is to be minimized.

$$\frac{1}{2} \|w\|^2 + C \sum_{j=1}^u \xi_j \quad (4)$$

The first term in (3) corresponds to the size of the margin and the second term represents misclassification. By solving a quadratic programming problem, the decision function  $f(x) = \text{sign}(g(x))$  can be derived where

$$g(x) = \left( \sum_{i=1}^n \lambda_i y_i (x_i \cdot x) + b \right). \quad (5)$$

The decision function depends on only support vectors ( $x_i$ ). Training examples, except for support vectors, have no influence on the decision function. Non-linear decision surfaces can be realized by replacing the inner product of (4) with a kernel function  $K(x, x_i)$  [14,18,19].

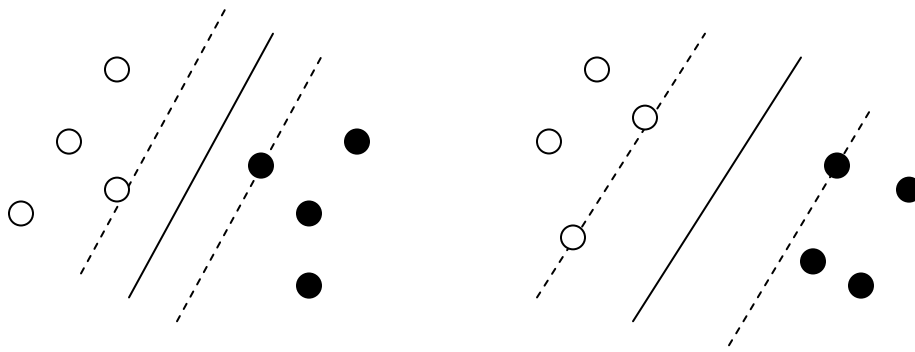
$$g(x) = \left( \sum_{i=1}^n \lambda_i y_i K(x_i, x) + b \right). \quad (6)$$

SVMs have advantage over conventional statistical learning algorithms from the following aspects:

1. SVMs have high generalization performance independent of dimension of feature vectors [15,20,21].
2. SVMs can carry out their learning with all combinations of given features without increasing computational complexity by introducing the Kernel function [15]
3. SVMs use a large margin to prevent over fitting, which does not necessarily depend on the number of features, as SVM has the potential to handle these large feature spaces [15].
4. The theoretical analysis concludes that SVMs acknowledge the particular properties of text: (a) high dimensional feature spaces, (b) few irrelevant features (dense concept vector), and (c) sparse instance vectors [11].
5. The experimental results show that SVMs consistently achieve good performance on text categorization tasks, outperforming existing methods substantially and significantly [2,21,22].
6. Another advantage of SVMs over the conventional methods is their robustness. SVMs show good performance in all experiments, avoiding catastrophic failure, as observed with the conventional methods on some tasks[2].
7. SVMs are the most accurate classifier and fastest to train [23,24].

### 4. System Architecture

The major components of the semantic document classification system includes the parser and stop words elimination module, feature extraction module, SVM classification system, ontology mapping module and the domain ontologies along with the other necessary features like the input, training and evaluation modules as shown in figure 2.



**Figure 1.** Two possible separating hyper planes

#### 4.1 Parser and Stop word elimination module

The text classification is carried out by transforming documents, which typically are strings of characters, into representations suitable for the learning algorithm and the classification task. The huge number of features in each hypertext document has to be reduced before being processed by classifiers as many of these features represent noise or irrelevant contents [25]. In case of semi structured documents the tags are to be removed such that classification is performed [1]. This function is carried out in this parser module. This step is followed by the stop word elimination process in which the most frequent words such as ‘to’, ‘and’, ‘it’, etc are removed to save spaces for storing document contents and reduce time for the search process.

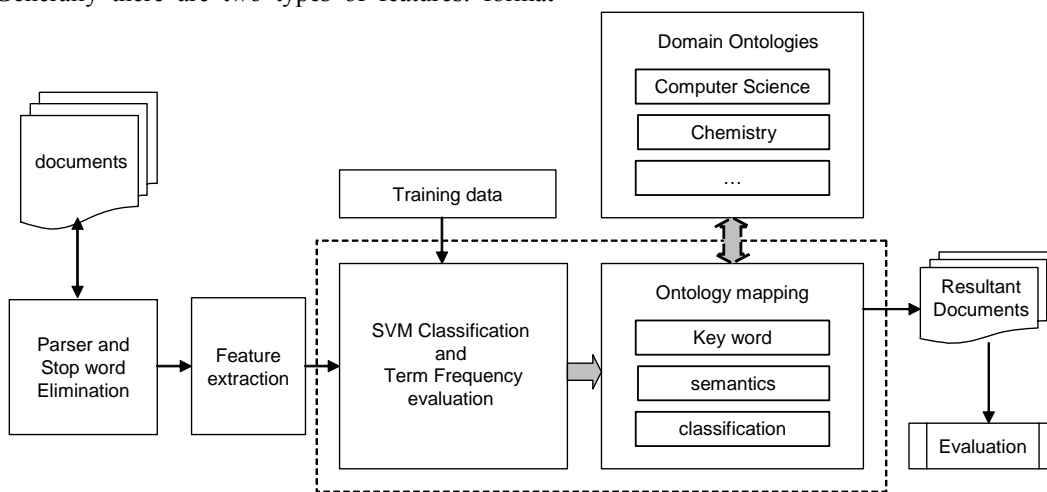
#### 4.2 Feature Extraction module

The work of extracting features from document is important for learning process[26].

Generally there are two types of features: format

features and linguistic features [27]. This feature Extraction module considers both format and linguistic features for extracting feature vectors. In the format feature extraction, there are four binary features that represent the normalized font size of the unit whether or not the current unit is in boldface. If the font size of the unit is the largest in the document, then the first feature will be 1, otherwise 0. If the font size is the smallest in the document, then the fourth feature will be 1, otherwise 0. If the font size is above the average font size and not the largest in the document, then the second feature will be 1, otherwise 0. If the font size is below the average font size and not the smallest, the third feature will be 1, otherwise 0. It is necessary to conduct normalization on font sizes. The linguistic features are based on key words and title to categorize them into a positive or a negative word. This binary feature represents whether or not the current word is one of the positive words or of the negative words.

These linguistic features are language dependent. By considering these two features, feature vectors for



**Figure 2.** Architecture of Semantic Document Classification System

the words are generated using the internal resource features such as orthography and external resource features such as matching against dictionaries.

### 4.3 SVM classification module

Each word obtained from the previous modules corresponds to a distinct feature vector and forms as the input for the SVM algorithm. SVMs are very well suited for text categorization. The experimental results show that SVMs consistently achieve good performance on text categorization tasks, outperforming existing methods substantially and significantly. With their ability to generalize well in high dimensional feature spaces, SVMs eliminate the need for feature selection, making the application of text categorization considerably easier [2]. SVM training is carried out with LIBSVM package [28] as it has higher accuracy, efficient, fast and mainly provides multi class classification [26]. The SVM performs a domain level classification of the words. The words are classified with respect to the domain using the domain knowledge in the ontology. The elimination of non informative words is performed [1]. The words are classified as positive or negative by the SVM. The domain related words are classified positive and the other words as negative. The words, which are classified positive, are interpreted from the original document and these words are considered as key words. Further, the term frequency of each generated keywords of the document is counted. A new data base is created in which the term frequency for the generated keywords is stored along with the key word.

### 4.4 Ontology mapping module

The key word that has maximum term frequency is extracted and is used in document classification. The mapping module will take the output of the extraction module as an input. The domain of the keyword is identified. The keywords will be mapped with the corresponding ontology concepts to which the key word is associated. The documents are classified if there exists a positive mapping. Else the keyword with next high term frequency is considered for classification using the mapping process. The domain of the document is first identified. The key word mapping will further lead to document classification at multi-levels with respect to the domain. During this mapping process the presence of semantic component instances for the key words are also checked. The semantic component instance is one or more segments of text that contains information about a particular aspect (semantic component) of the concept instance (a topic) that is the document focus [29]. Although

documents are often organized around semantic components, they do not always correspond to structural elements, such as subheadings, and a single semantic component instance may be composed of multiple discontinuous segments of text [30]. The identification of semantic components is an automatic process performed in association with the ontology mapping of key words for which the domain knowledge is made available in the form of ontologies. The documents may be classified under multiple categories in the mapping module. The semantic components are useful for making a precise classification. By identifying the semantic components the documents relevance to a category can be identified and thus the document can be appropriately categorized.

### 4.5 Domain Ontologies

The domain ontology for the computer science domain is extracted from the Dmoz Directory. The chemical ontologies comprising of compound ontology, reagent ontology and reaction ontology are used from the Chemical Ontological Support System (COSS) developed earlier for the representation of chemical reactions and their mechanisms [31]. The domain knowledge present in the ontology is used in the first level of classification of words with respect to the domain using SVM. In the second level the document is classified based on the key word with maximum term frequency using the domain knowledge. The knowledge is further used at the third level for identifying the semantic components in the document which provides a precise classification. Again the same domain ontologies are used to fix the exact category of the resultant document. This is feasible because the contents are stored in the ontologies in a hierarchical fashion; top priority term automatically fixes its hierarchy from the ontological concepts. This allows the possibility of classifying the documents under multiple categories too. There by the use of domain ontologies at multiple levels provide a refined and semantically rich classification output. The system is easily extensible simply by including more and more domain ontologies of different disciplines appropriately.

## 5. Implementation

The implementation of the system is done in windows platform using JAVA. The SVM classification is done using LIBSVM which can perform multi class classification. Ontology for computer domain is constructed from DMOZ directory

and for the chemical ontology the COSS developed in XML is converted into text format and used. The document classification is done for the documents related to Computer Science and Chemistry domains. The documents were classified with good accuracy. The documents related to these two domains when fed into the system, the domain of the document was

precisely identified and the documents were classified at multi levels with respect to the domain. This system can be extended to other domains with minimal input of knowledge about the domain. The snapshots of output of the various modules from the computer science and chemistry domains are given below.

### Input document

#### Computer Science Domain

In the coming years a very real and very near future will be emerging. Hollywood may have not exaggerated the change. **expensive, extremely secure technology.** If you have ever been to a meeting, you already know **Biometrics** can be truth to this claim, depending on how biometric systems are implemented. For example, a retina scanner provides little security if it is the door for a stranger standing behind them. **Biometrics** can only provide effective security when properly combined with other identification

#### ChemistryDomain

are broadly grouped by structure and property differences. **aromatic** compounds that contain aromatic rings are further grouped into **alkanes** that contain single bonded C-C structures, **alkynes** with triple bonded C≡C structures, **cycloalkanes**, or ringed aliphatic compounds. Each **functional** groups because they give the characteristic reactivity and fate. **1. Structure and physical properties**

### Output of the parser, stop word elimination module

identification technologies. technologies use individual's iris patterns, facial characteristics. biometrics appropriate technical ability catalogue track them. traits easier to store, example, relatively simple record store database. improving accuracy security measurements reducing cost. fingerprint readers, become reliable. accuracy systems retina scanner, digital security ever-growing demand. Complex to implement, response call effective security. Biometrics, Past Present and future futuristic technology science fiction movies, give a reputation. business owners executives likely say, "I've convinced it's realm science fiction. result, biometric systems are emerging. Taking Life Measurements Illustration Figure 1. marketed advanced, high-end security technology years ago. fact, movies soon front door home. Door locks work in real market biometrics emerging. Hollywood exaggeratedly secure technology. pass retina scanner meeting, depending biometric systems implemented. example

reaction, without input light heat energy. Inhibitors slow reaction proceeds normally. These reactions occur sewage production low level concentrations halomethanes treatment. 4. Kinetics activation energy Reaction kinetics measure rate thermodynamics completely independent approaches chemical thermodynamics kinetics time. Energy Activation energy of reactant molecules successfully conclude reaction resulting insufficient kinetic energy bounce off, react. reactant molecule successful collisions increases, rate increases. Page 3 Example greater effect changing temperature. relationship temperature constant = preexponential function e = natural log Ea = a state. transition state condition reaction top activation energy. hill. energy activation thought enthalpy formation intermediate structure transition state complex, useful concepts examine HCl ΔH=1kcal/mol Ea = 4 kcal Products Reactants Page 4 structure, properties follow pattern laid methane. 1. Structure 2 carbons, formula C<sub>2</sub>H<sub>6</sub> carbon bonded 4 atoms, et

### Output of the SVM module (interpreted)

Security Storage Biometrics Life security biometrics organizations biometrics Biometric Recognition systems database systems Bio authentication system access devices system scanners PC computers fingerprint used at recognition hardware processing hardware images face face recognition face recognition biometrics systems Systems Biometrics pas

Hydrocarbons Aliphatic Hydrocarbons Hydrocarbons aliphatic aromatic aromatic aliphatic aliphatic groups functional groups compound electron water atom product water water water water alkane Alkyl groups alkane compound alkyl atom Alkane Alkane alkyl alkyl groups alkyl groups alkane primary atom atom atom alkane organic water alkane

### Output of the term frequency evaluation(keywords identified)

Biometrics  
 security  
 systems  
 recognition  
 Face  
 technologies  
 fingerprint  
 people

alkane  
 Alkyl  
 Hydrocarbon  
 Aliphatic  
 water  
 atom  
 saturated  
 organic

### Output of the mapping module – initial classification

### Computer Science domain

```
THE KEYCOUNT IS 1
THE KEYWORD IS Biometrics
COMPUTER DOCUMENT
result computerscience:Security:Biometrics
```

### Chemistry Domain

```
THE KEYCOUNT IS 1
THE KEYWORD IS alkane
CHEMISTRY DOCUMENT
result organicCompound:hydrocarbon:aliphaticHydrocarbon:aliphaticSaturatedHydrocarbon:alkane
```

### Output of the semantic components identified

```
Authentication
Biometrics
organizations
Face
recognition
fingerprint
iris
software
```

### Output of the final classification (along with the frequency of occurrence)

#### Computer Science Document

```
computerscience:Security:Biometrics:Face Recognition 9
computerscience:Security:Biometrics:Fingerprint Recognition 8
computerscience:Security:Biometrics:Iris Recognition 2
computerscience:Security:Biometrics:Organizations 1
computerscience:Security:Biometrics:Software 5
```

#### Chemistry Document

```
organicCompound:hydrocarbon:aliphaticHydrocarbon:aliphaticSaturatedHydrocarbon:alkane
```

Based on the frequency of the semantic components, the documents relevance to that category can be identified. Thus the document was precisely classified. But some documents will be categorized under several categories in the ontology mapping module. The document related to algorithms was classified by this system under 2 categories.

```
| computerscience:Algorithms
computerscience:Artificial_Intelligence:Genetic_Programming:Algorithms
```

The semantic components were identified from the documents with the aid of knowledge present in the ontology and they are

```
|Algorithms
numbers
Sorting
Searching
```

The document was precisely classified based on these semantic components.

```
computerscience:Algorithms:PseudorandomNumbers 1
|computerscience:Algorithms:Sorting 7
computerscience:Algorithms:SortingandSearching 6
```

## 6. Results

To evaluate the effectiveness of this approach, a comprehensive performance study has been conducted with web documents as input. The performance of text categorization using SVM with the aid of domain ontology is measured using the following performance measures. Assuming binary classification (relevant/not relevant), the following performance measures are evaluated.

a = the number of relevant documents, classified as relevant  
 b = the number of relevant documents, classified as not relevant  
 c = the number of not relevant documents, classified as not relevant  
 d = the number of not relevant documents, classified as relevant.  
 Obviously, the total number of documents N is equal to  $N = a + b + c + d$

The performance measures such as precision, recall, F-measure can be defined as follows

Precision: Number of correctly identified items as percentage of number of items identified.

$$\text{Precision} = \frac{a}{a + d}$$

Recall: Number of correctly identified items as percentage of the total number of correct items.

$$\text{Recall} = \frac{a}{a + b}$$

F-measure: Weighted Average of precision and recall.

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The performances of Naïve-Bayes, simple SVM, SVM integrated with domain ontology is compared with this system (minimal semantic components identified) and the graph was plotted for different training set size. The performance of NB classification is the least, like expected. SVM result has reasonable F-measure, while SVM with the ontology mapping outperforms both but has less performance compared to this system. The use of ontology and semantic components makes the classification more accurate. The training size has a

very high impact on the performance of the Naïve Bayes and SVM classification methods. These methods provide a good, consistent F-measure only when the training size of the documents goes beyond 100. In this system as shown in the comparative evaluation graph in figure 3, the training size of the documents does not have much influence on the performance of the classification. This clearly shows that the training size need not be more for high F-measure which is a significant advantage of the proposed system.

To improve the classification performance the semantic components were identified. Some of the documents were classified under multiple categories in the ontology mapping module. But it has been found that the documents are categorized under some irrelevant categories. The advantage of semantic component identification as shown in figure 4 is that the documents are being precisely classified under relevant categories. This greatly reduces the irrelevant categorization of the documents which is another advantage of this system.

The identification of the semantic components further increases the accuracy of the system. Although the SVM based classification using ontology had a high F-measure compared to other methods the accuracy is further increased when the knowledge of semantic components was used. The classification accuracy increases when more number of semantic components is identified as shown in Figure 5.

## 7. Conclusion

The presented study describes the usefulness of integrating the domain ontologies and the associated semantics with the conventional SVM classification to obtain an automated document classification resulting

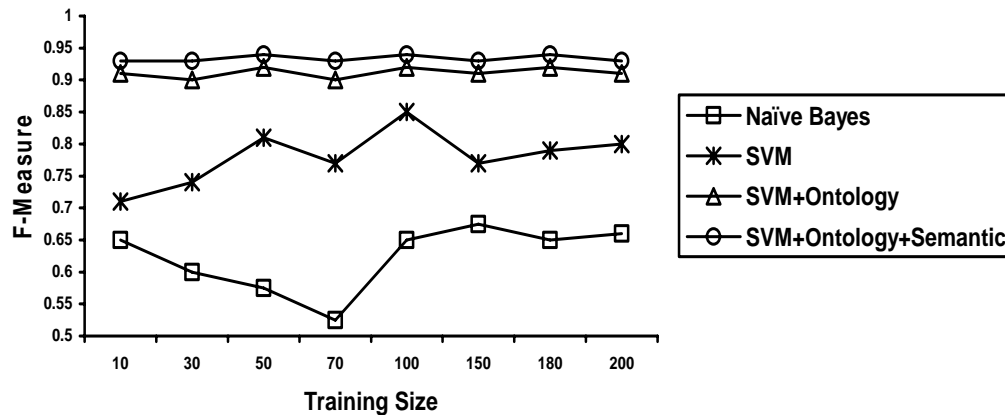
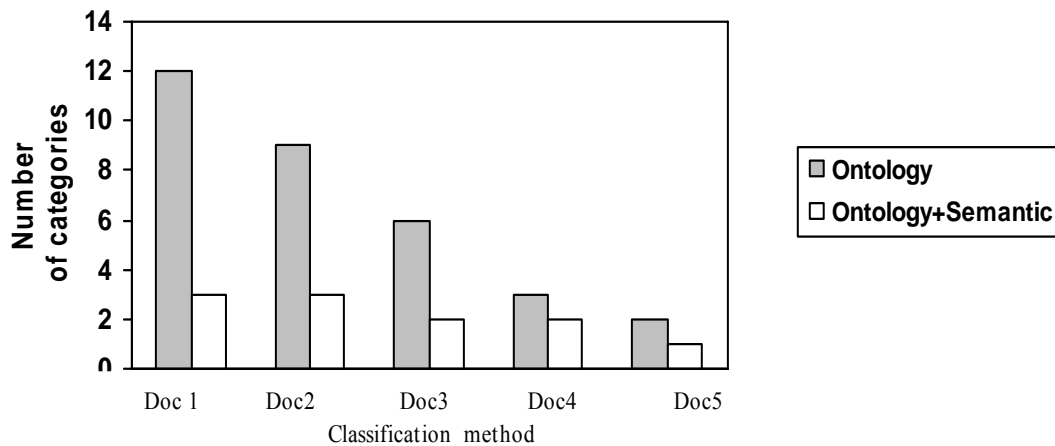


Figure 3. Comparative Evaluation



**Figure 4.** Advantage of semantic component identification

in a precise and meaningful classification. This work is distinguished from other studies in the following areas:

1. The key words that are required for classification of the document are classified using SVM.
2. The Term Frequency of these key words alone is found which reduces the processing time by a considerable amount.
3. The training size of the documents is very less and hence training time is greatly reduced.
4. The feature space dimension is also very less.
5. Multi level classification of the documents is done using domain ontology which is more accurate.
6. The semantic components are identified by an automated method.
7. Describing the content of the documents in domain-specific collections, using *document classes* and *semantic components*, will supplement existing indexing and searching techniques and improves information retrieval. Semantic components eventually will be a useful supplement to other types of searching and will facilitate more precise retrieval of domain-specific documents.
8. Identification of semantic components provides convenient and effective ways for users to annotate Web pages with RDF metadata, thus facilitating wider availability of semantic Web content.

All these advantages make this system a promising and efficient method for document classification. Future work includes testing the usefulness of semantic components to searchers and testing the accuracy and consistency of identifying semantic components.

## 8. Acknowledgement

This research work was supported by the Research Grant from All India Council for Technical Education(AICTE), New Delhi, India under Research Promotion Scheme(RPS) 2004-2006 F.No.8022/RID/NPROJ/RPS-109/2003-04.

## 9. References

- [1] Wang,Z.,Sun,X., And Zhang,D, "An optimal text categorization Algorithm based on SVM," Proceedings of the International conference on communications,circuitsandsystems,(June2006),Volume: 3, 2137-2140.
- [2] Joachims, T, "Text categorization with supportvector machines: Learning with many relevant features," Proceedings of ECML-98, 10th European Conference on Machine Learning, 1997.
- [3] Wu,S., Tsai,T.,And Hsu,W, "Text categorization using automatically acquired domain ontology," Proceedings of the sixth international workshop on Information retrieval with Asian languages, volume 11, 138-145, 2003.
- [4] Song,M., Lim,S., Kang,D., And Lee,S, "Automatic classification of web pages based on the concept of domain ontology," Proceedings of the 12th Asia-Pacific Software Engineering Conference (APSEC'05), 15-17 (Dec. 2005).
- [5] Price,S., Delcambre,L.,Nielsen,M.L., Tolle,T., Luk,V., And Weaver,M, "Using semantic components to facilitate access to domain-specific documents in Government settings," Proceedings of the 2006 international conference on Digital government research, vol.151, 25-26, 2006.
- [6] Fuhr, N., Hartmann, S., Lustig, G., Schwantner, M., And Tzeras, K, "Air/x—A rule-based multistage indexing systems for large subject fields", In: Proceedings of RIAO'91. 606–623,1991.
- [7] Creecy, R..H., Masand. B.M., Smith, S.J., And Waltz,

D.L., "Trading mips and memory for knowledge engineering: Classifying census returns on the connection machine," *Comm. ACM.* 35:48–63, 1992.

[8] Tzeras, K., And Hartman, S, "Automatic indexing based on bayesian inference networks," In: *Proc. 16<sup>th</sup> Ann. Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*,. 22–34.

[9] Lewis, D.D., And Ringuette, M., "Comparison of two learning algorithms for text categorization," In: *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94)*, 1994.

[10] Apte, C., Damerau, F., And Weiss, S, "Towards language independent automated learning of text categorization models," In: *Proceedings of the 17th Annual ACM/SIGIR Conference*,1994.

[11] Wiener,E., Pedersen, J.O., And Weigend, A.S, "A neural network approach to topic spotting," In: *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95)*.

[12] Cohen, W.W., And Singer, Y, "Context-sensitive learning methods for text categorization," In: *SIGIR '96: Proceedings of the 19th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 307–315, 1996.

[13] Yang,Y, "An evaluation of statistical approaches to text categorization," *Published in Information Retrieval journal.* volume 1,Numbers 1-2, April 1999.

[14] Hirao,T., Isozaki,H., And Maeda,E, "Extracting important sentences with support vector machines," *Proceedings of the 19<sup>th</sup> international conference on computational linguistics*, volume 1, 1-7, 2002.

[15] Kudoh,T. And Matsumoto,Y, "Use of support vector learning for chunk identification," In: *Proceedings of CoNLL-2000 and LLL-2000*,142-144. Lisbon, Portugal, 2000.

[16] Chen,C., Lee,H., And Kao,M, "Multi-class SVM with negative data selection for web page classification," *Proceedings of IEEE International Joint conference on Neural Network*, 2004.

[17] Lam,W., Ruiz,M., And Srinivasan,P, "Automatic text categorization and its application to text retrieval," *IEEE transactions on knowledge and data engineering.* vol. 11,issue no. 6, 865-879, 1999.

[18] Tong,S., And Koller,D, "Support Vector Machine active learning with applications to text classification," *The Journal of machine learning research.* volume 2, (March 2002), 45-66.

[19] Sato,K., And Saito,H, "Extracting word sequence correspondences with SupportVectorMachines," *Proceedings*

of the 19th international conference on Computational linguistics, Volume 1, 1-7, 2002.

[20] Sebastiani,F, "Machine learning in automated text categorization," *ACM Computing Surveys.* Vol. 34, No. 1, 1–47, 2002.

[21] Lee,C., And Yang,H, "A classifier-based text mining approach for evaluating semantic relatedness using Support Vector Machines," *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05)*, Volume 1, Issue , 4-6 April 2005 ,128 – 133, 2005.

[22] Yang,Y., And Liu,X, "A reexamination of text categorization methods," *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, 1999.

[23] Dumais,S., Platt,J., Heckerman,D., And Sahami,M, "Inductive learning algorithms and representations for text categorization," *Proceedings of the seventh international conference on Information and knowledge management*, 1998.

[24] Cai,L., And Hofmann,T, "Hierarchical document categorization with Support Vector Machines," *Proceedings of the thirteenth ACM international conference on Information and knowledge management (CIKM 2004)*, 78-87, 2004.

[25] Moradi,P., Abdollahzadeh,A., And Ibrahim Shiri,M, "Novel method for improving web text classifiers performance through machine learning," *2<sup>nd</sup> Information and Communication Technologies( ICTTA '06)*.volume 1,(April 2006), 534-539, 2006.

[26] Jun,H., And Houkuan,H, "An algorithm for text categorization with SVM," *Proceedings of IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering (TENCON'02)*.Volume 1, 47 – 50, 2002.

[27] Hu,Y., Li,H., Cao,Y., Meyerzon,D., And Zheng,Q, "Automatic extraction of titles from general documents using machine learning," *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, 145-154, 2005.

[28] Chang,C., And Lin,C, "LIBSVM- A library for Support Vector Machines," <http://www.csie.ntu.edu.tw/~cjlin/libSVM>

[29] Price,S.L, "SEMANTIC COMPONENTS: A new model for enhancing retrieval of domain specific information," *10<sup>th</sup> ECDL 2006*.

[30] Price,S.L.,Delcambre,L.M., And Nielson,M.L, "Using semantic components to express clinical questions against document collections," *Proceedings of the international workshop on Healthcare information and knowledge management (HIKM '06)*. 9-16.

[31] Sankar,P., And Aghila,G, “Design and Development of Chemical Ontologies for Reaction Representation,” Journal of Chemical Information and Modeling, Volume 46(6), 2355-2368, 2006