Dynamic Ontology Based Model for Text Classification

K. Purna Chand
Research Scholar, Department of Computer Science and Engineering, JNTU College of Engineering, Kakinada, Andhra Pradesh, India.

Dr. G. Narasimha
Associate Professor, Department of Computer Science and Engineering, JNTUH College of Engineering, Jagitial, Telangana State, India.

Abstract
Text Mining is the process of extraction of knowledge from large amounts of text. Information retrieval gains more focus due to tremendous increase of textual information in web servers. It is a very challenging task to process, organize, analyze and extract knowledge from huge volumes of unstructured text. Traditional text classification algorithms need well defined text corpus to train and classify the given textual information. It is a very complex task to build text corpus with the help of thesaurus on the given text documents. To overcome this, ontology models can be considered for constructing knowledge base. Domain ontologies are considered for extracting more useful and high quality data on a particular domain. Many traditional approaches use static methods to construct domain ontologies. But, ontologies developed by these static approaches consist of limited terms in their knowledge base due to lack of updation. In this paper, we propose a dynamic ontology based model to classify the extracted terms and to build knowledge base on a particular domain. Experimental results show that our dynamic ontology based model performing excellently by frequent updating of extracted terms in our knowledgebase.

Keywords: Text Mining, Information retrieval, Dynamic ontology model, Domain ontologies, Knowledgebase

Introduction
Now a day’s many organizations allows users to store their data mostly in structure independent format. Web servers contain volumes of massive data in unstructured and/or semi structured format. It is very difficult to process, organize, analyze and extract knowledge from huge volumes of unstructured text [8][9]. Information retrieval systems are needed to be automated and more accurate to provide qualitative results for any decision making process. Text Mining is the process of extraction of knowledge from huge volumes of unstructured text. Text mining integrates text classification and text document clustering. Text classification is the process of developing a classifier on given text documents to identify class labels for distribution. And text clustering will identifies the similar search patterns from unstructured text. Text classification techniques are used in many applications, including e-mail filtering, spam filtering, news monitoring, sorting through digitized paper archives, automated indexing of scientific articles, classification of news, stories and searching for interesting information on the web, biomedical applications [7]. However, it is very difficult to get well-classified corpus. Even if any corpus is available it may be classified improperly due to the less terms classified by the limited and static nature of classifiers. To overcome this, we propose an ontology based model to classify the terms and prepare the knowledge base. Ontology is a data model which represents knowledge in terms of set of classes or concepts and relationships among them. The classes define the types of attributes, or properties common to individual objects within the class. “Seth et.al 2002; Moreover, classes are interconnected by relationships, indicating their semantic relations” [1]. Ontology is a data model which represents knowledge in terms of set of concepts and relationships among them. The main objective of Ontologies is knowledge management which is being done by ontology languages like RDF and OWL [9]. RDF provides syntax for data and metadata representation. It consists of set of statements and data models. RDF statements can be formed from a triplet (Subject, Property and Object). RDF data model is a directed graph, nodes represent subjects or objects and arcs represent the properties. Nodes are labeled as URIs to define resources or literals (i. e. strings or numbers). Nodes which are unlabeled, called blank nodes. Blank nodes are commonly used to group properties. Edges are always labeled by URIs representing relationships between the subjects and the objects. Specific applications of vocabularies can be created by a RDF Schema (RDFS). It extends RDF by defining the schema of RDF statements. RDFS will guide Ontologies to understand the relationships between RDF items. RDF and RDFS describe knowledge about a specific domain with a set of instances. OWL works on the descriptions of the objective world from two aspects: the concepts and the attributes. The description methods are the object-oriented methods which use RDFS and OWL’s own grammars for describing concepts, inherited relations of classification as well as associated relationship. OWL supports XMLS’ all data types for describing and expressing the concepts and attributes. OWL has the ability to describe complex semantic web and semantic inference rules.

Related Work
Traditional classification algorithms use statistical approaches or machine learning methods to perform the classification. Such methods include Naive Bayes [2], Support Vector
Machines [18], Latent Semantic Analysis [14] and many others. These methods require a training set of pre-classified documents that are used for classifier training [4] later, the classifier can correctly assign categories to other, previously unseen documents.

During the last decades, many approaches have been proposed machine learning algorithms for supervised and unsupervised text categorization. However, existing text categorization systems have typically used the Bag of Words model where single words or word stems are used as features for representing document content [5].

However, in recent years’ current researches working rigorously to incorporate semantic background knowledge into text classification. WordNet is a big lexical database of English. All grammatical basics like Nouns, adjectives, verbs and adverbs are grouped into set of synonyms called as synset, each synset states about a distinct concept. WordNet resembles a dictionary, which groups’ words based on their meanings. The words that are in close proximity to one another in the network are semantically clarified. Latent Semantic Analysis [16] offers an attractive way to transition from the word-space to the concept-space of related phrases. The extracted concepts can be effectively applied to classical categorization, as presented in [19].

There exist few ontology-based approaches. Ontologies [15] offer knowledge that is organized in a more structural and semantic way. The knowledge represented in a comprehensive ontology can be used to identify concepts in a text. Furthermore, if the concepts in the ontology are organized into hierarchies of higher-level categories, it should be possible to identify the category that best classify the content of the text.

Ontology can be successfully used for term disambiguating and vocabulary unification, as presented in [13]. Another approach, presented in [11] reinforces occurrence of certain pairs of words or entities in the term vector that are related in the ontology.

The significance of our approach is that, it does not require well-defined text corpus to train the classifier. Our proposed approach will find the semantic relations between classified text and ontology terms. And also dynamically updates the knowledge-base for every newly added term.

**Dynamic Ontology Model**

“Neda Ghiassi [17] research describes ontologies are mostly used to specify the knowledgable domains. Ontology is a data model which represents knowledge in terms of set of concepts and relationships among them”. The main objective of Ontologies is knowledge management it can be done by ontology languages like RDF and OWL.

Domain Ontology provides vocabulary to represent conceptual structure of a particular domain. They are usually connected to some top-level ontology so that there is no requirement to include the common basic knowledge.

Domain ontology model considers only the set of entities, their relationships, and the taxonomy of categories from the given text documents. In these approaches the ontology effectively becomes the classifier. The proposed approach completely transforms the document text into ontological structure.

Our proposed dynamic ontology model can be explained by the following modules.

- **Document Analysis**
- **Ontology Construction**

![Architecture of Dynamic Ontology Model](image_url)

**Figure 1: Architecture of Dynamic Ontology Model**

**Document Analysis**

This module defines the total process of Term extraction i.e. taking unstructured text documents for analysis to construction of knowledge-base. It describes about two functionalities.

- **Term Processing**
- **Thesaurus Generation**

**Term Preprocessing**

The functionality of this module is to process unstructured text by retrieving only the relevant documents and analyzes their word’s relative frequencies with the help of statistical measures. The relevant documents can be assessed by well-known measures like Precision (PC), Recall (RC) and F-Measure (FM).

\[
\text{PC} = \frac{\text{Number of correct items which are classified}}{\text{Number of items which are classified}} \\
\text{RC} = \frac{\text{Number of correct items which are selected}}{\text{Number of selected items which are correctly classified}} \\
\text{FM} = \frac{\text{Harmonic mean of precision and recall}}{\text{FM}}
\]

After extracting the relevant documents we applied the Stemming and Stop-Word removal procedures to reduce the number of term occurrences by considering only most useful terms.

**Thesaurus Generation**

Thesaurus is a bag of words grouped together according to similarities of their meaning, in contrast a dictionary, which provides definitions for words, and generally lists them in alphabetical order. The main objective of thesaurus is to find the words by which we can quickly express the idea.
Although including synonyms, a thesaurus should not be taken as a complete list of all the synonyms for a particular word. The entries are also designed for drawing distinctions between similar words and assisting in choosing exactly the right word. Unlike a dictionary, a thesaurus entry does not give the definition of words.

The detail description of our work is explained below. Here, we use Wordnet; a lexical semantic analyzer as a thesaurus to develop our knowledge-base.

**Wordnet**

WordNet® is a big lexical database of English. It generates word senses as synsets. Synset defines a group of parts of speech like nouns, adjectives, verbs and adverbs. Each synset states about distinct concepts. Wordnet is very similar to a dictionary, it consists of words collected based on their meanings. The main objective of Wordnet is to define the semantic relations between the words in all the senses and to identify the words which are in close proximity to one another in a text document.

**Knowledgebase**

A knowledgebase (KB) is a technology used store the information retrieved from structured or unstructured documents. The main purpose of knowledge-base is to represent facts about the world. The Knowledgebase can be constructed by our previous work Semantically Enriched Terms (SET) Clustering Algorithm. This extracts the most similar terms from the given set of text documents and constructs the knowledgebase. But, the entire process is static. So, in view of this, we propose an Improved Semantically Enriched Terms (ISET) clustering algorithm, in this knowledgebase will be updated dynamically for every occurrence of new terms extracted from the new documents.

**Improved Semantically Enriched Terms (ISET) Clustering Algorithm**

The main objective of this algorithm is to construct a Knowledgebase as hierarchical tree structure. Root node represents the domain; intermediate nodes represent class labels and all leaf nodes represents the semantically related terms extracted from given documents.

After Stemming and Stop word removal processes, our proposed ISET clustering algorithm as shown in Fig. 2 will work. We considered Wordnet 3.0 as thesaurus to identify synonyms as Synset and polysemies as Polyset (P_Set). Knowledgebase is represented like a tree structure and it was constructed by using Synset (S_Set). Finally, dynamic ontology model has been developed from the updated knowledgebase. For the occurrence of every new term in the test dataset, we sent it to a New_Term_Set (NT_Set). For all the terms in NT_Set will be immediately updates our knowledge base. Finally, ontology will be constructed dynamically according to the new terms.

---

**Algorithm ISET_Clustering (D, T, K)**

// D-Document Set. Consisting of ‘n’ no of documents. Ex: D= {d1, d2,..., dn}.
// T-Thesaurus. Here, we use Wordnet 3.0; a lexical semantic analyzer as a thesaurus to develop our knowledge-base.
// K-Knowledgebase. Here knowledgebase constructed like a tree structure.

**Step 1:** Given a collection of text documents \( D = \{d_1, d_2,..., d_n\} \).

*Applied Stemming and Stop-Word Removal algorithms. Consider the remaining terms as Unigram(Ug)-Frequently Occurring 1 Word Bigram(Bg)-Frequently Occurring 2 Words Trigram(Tg)-Frequently Occurring 3 Words Multigrams(Mg)-Frequently Occurring 4 or more Words.

**Step 2:** Assign ranks to each term based upon their relative frequencies in a single document or in clustered documents.

\[ \text{Rank} = \text{Term Frequency (TF)} \times \text{Min Support} = 2 \]

**Step 3:** Identify the semantic relationship between the terms by using a Lexical Semantic Analyzer Wordnet 3.0

\[ \text{Sem}_\text{Rel(Terms)} = \text{Synonyms or Estimated Relative Frequency} \]

**Step 4:** The semantically enriched terms are grouped into different categories. According to their similarities different Synset(S_set) will be formed.

\[ \text{Synset (S_Set)} = \{\text{Class label (l)}, \text{Sem}_\text{Rel(Terms)}\} \]

**Step 5:** Tag the class_labels (l) for the extracted terms using term-concept relationship.

**Step 6:** Construct the Knowledge_base (KB) as tree like structure for the extracted class_labels (l).

**Step 7:** For every new text dataset repeat the steps 1 to 5. Identifies the occurrence of new terms and group them as New_Term_Set(NT_Set).

**Step 8:** For every newly occurring term in New_Term_Set (NT_Set) update the Knowledge_base (KB)

**Step 9:** The updated Knowledge_base (KB) will be used to build Ontologies dynamically.

---

Figure 2: Improved Semantically Enriched Clustering Algorithm.
Experimental Setup
The accuracy of our dynamic ontology model is compared to the static model and observed the performance improvements. The algorithms presented in this paper make use of the domain ontology described in the previous section. The experiments have been done on the wiki data set. We considered three test datasets containing the 500 words in each.

We calculated the classifier building times in milliseconds and classification accuracy for each test dataset. The following results are occurred when calculated with the Weka tool. We have calculated the mean classification time and accuracy of both statistical and dynamic ontology models. And, we concluded that dynamic ontology model is outperforming by taking less time and high accuracy.

Table 1: Experimental Outcomes on Static and Dynamic Ontologies.

<table>
<thead>
<tr>
<th>Test Data Sets</th>
<th>Static Ontology</th>
<th>Dynamic Ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (ms)</td>
<td>Accuracy</td>
<td>Time (ms)</td>
</tr>
<tr>
<td>Dataset-1</td>
<td>6.7</td>
<td>0.86</td>
</tr>
<tr>
<td>Dataset-2</td>
<td>7.4</td>
<td>0.88</td>
</tr>
<tr>
<td>Dataset-3</td>
<td>8.2</td>
<td>0.91</td>
</tr>
<tr>
<td>Mean</td>
<td>7.43</td>
<td>0.88</td>
</tr>
</tbody>
</table>

The performance analyses of three datasets are shown in terms of a graphical representation and shown as following. Data labels in the plot are presenting that our dynamic ontology took less classification time and more accuracy.

Figure 3: Performance Analysis of Static and Dynamic Ontologies

Now, we examined the performance of our dynamic ontology model by integrating with Improved Semantically Enriched Clustering (ISET) algorithm. The performance of our dynamic ontology model with updated Knowledgebase is compared with traditional classifiers, we observed that our ISET Ontology model classifying more terms in very less amount of classification times and the classification accuracy is also improved.

We observe the performance improvements of our proposed ISET-dynamic ontology with the traditional classifiers as tree based classifiers like Decision Tree (ID3), Probabilistic classifier like Naïve-Bays and support vectors based SVM.

Table 2: Comparison of ISET-Ontology model with traditional classifiers

<table>
<thead>
<tr>
<th>Measure</th>
<th>ID3</th>
<th>Naive Bays</th>
<th>SVM Classifier</th>
<th>ISET-Ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.221</td>
<td>0.443</td>
<td>0.598</td>
<td>0.684</td>
</tr>
<tr>
<td>Recall</td>
<td>0.426</td>
<td>0.562</td>
<td>0.672</td>
<td>0.786</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.536</td>
<td>0.648</td>
<td>0.762</td>
<td>0.846</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.486</td>
<td>0.522</td>
<td>0.69</td>
<td>0.89</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.42</td>
<td>0.54</td>
<td>0.60</td>
<td>0.80</td>
</tr>
</tbody>
</table>

We applied all the text retrieval methods on the total data set and finally calculated the accuracy of each classifier. The results obtained are showing that our ISET-Ontology model performing better than the specified classifiers.

Figure 4: Performance Improvements of ISET-DynamicOntology Model.

CONCLUSION
In this paper, we propose a dynamic ontology based model to classify the extracted terms and to build knowledge base on a particular domain. The main purpose of this model is, it depends only on Ontologies for classification instead of traditional training data sets. Our work has been explained in two modules. In the first approach, we have analyzed all the given input text documents. In order to extract the terms, we used our previously proposed Semantically Enriched Terms (SET) clustering algorithm. SET Clustering used Wordnet as a thesaurus to find the term-concept relationships. A Knowledge_base(KB) can be constructed like a tree structure by having Class_label (l) as a root node and all the related terms as its children. For every new test dataset our Knowledge_base(KB) will be updated.

The second approach, explained about building of Ontologies by taking the conceptual terms and relationships from the knowledgebase with the help of OWL editors like protégé. For every occurrence of a new term our Knowledge_base (KB) will be updated and immediately those updates are applied on our ontology model.
The experimental results shown in this paper describes the performance improvement our proposed dynamic ontology model with SET clustering algorithm(SET-DOM) over traditional classifiers like ID3, Naïve Bayes, and SVM Classifiers. The efficiency of any algorithm can be measured by the factors like accuracy and time. Hence, our model proved to be efficient.

Finally, our model can be enhanced and improvised further by applying machine learning concepts to automate the knowledge_base (KB) generation with the help of our Improved SET clustering algorithm. We are also planning to use machine learning for auto-restructuring of Knowledge_base (KB) and immediate ontology updating.

References


