

# A Novel Approach for Improving Similarity Search Using SVM Classification Algorithm

Anitha S.<sup>1</sup>, Radha P.<sup>2</sup>

<sup>1</sup> Research Scholar, Department of Computer Science, Vellalar College for Women Tamilnadu, India

<sup>2</sup> Assistant Professor, Department of Computer Science, Vellalar College for Women Tamilnadu, India

\*\*\*

**Abstract** - *The World Wide Web has transcended from a read-only to a read-write web. The problem of identifying important online or real life events from large textual document streams that are freely available on the World Wide Web is increasingly gaining popularity, given the flourishing of the social web. Earlier work used efficient algorithm for detecting all important events from a document stream through named entity recognition and topic modelling. In existing scenario, the learning algorithms are considered to identify and detect the unknown important topics from the disaster document. The problem found in this work is that the effective recognition of interesting named entities in previously unknown free text documents remains an open problem. The problem of finding similar literals with different meaning, or different literals that describe the same entity should be addressed. Hence to overcome all these issues the Support Vector Machine (SVM) algorithm is utilized in the current work. The SVM is used to create a model which helps to identify most random and important events from the given document. It is also used to recognize the high similarity sentences and semantic events from the specified document. The required information is retrieved based on the ranking events. The top most ranked events will identify the important event as well as produce the rich semantic meaning. Ranking SVM can be successfully applied for the task of finding and learning a similarity function for the event identification problem.*

**Key Words:** *World Wide Web, Entity Recognition, Support Vector Machine, Semantic.*

## 1. INTRODUCTION

Data mining is a recently emerging field, connecting the real world Databases, Artificial Intelligence and Statistics. The information age has enabled many organizations to gather large volumes of data. However, the usefulness of this data is negligible if “meaningful information” or “knowledge” cannot be extracted from it. Data mining is also known as knowledge discovery in databases. Data Mining refers to extracting or mining interesting

knowledge from large amounts of data stored either in databases, or other information repositories

Data mining involves integration of techniques from multiple disciplines such as database technology, statistics, machine learning, neural networks, information retrieval, etc. Data mining is the process of discovering meaningful patterns and relationships that lie hidden within very large databases. Data mining is a part of a process called knowledge discovery in databases (KDD). This process consists basically of steps that are performed before carrying out data mining, such as data selection, data cleaning, pre-processing, and data transformation.

There are many other terms carrying a similar or slightly different meaning to data mining such as knowledge mining from databases, knowledge extraction, Data/pattern analysis, Data archaeology and Data dredging. A standard definition for data mining is the non-trivial extraction of implicit, previously unknown, and potentially useful knowledge from data.

### 1.1 Support Vector Machines (SVM)

Support Vector Machine (SVM) is based on statistical learning theory. SVMs were initially developed for binary classification but it could be efficiently extended for multiclass problems. The support vector machine classifier creates a hyper plane or multiple hyper planes in high dimensional space that is useful for classification, regression and other efficient tasks. SVM have many attractive features due to this it is gaining popularity and have promising empirical performance. SVM constructs a hyper plane in original input space to separate the data points. Some time it is difficult to perform separation of data points in original input space, so to make separation easier the original finite dimensional space mapped into new higher dimensional space. Kernel functions are used for non-linear mapping of training samples to high dimensional space. Various kernel function such as polynomial, Gaussian, sigmoid *etc.*, are used for this purpose. SVM works on the principal that data points are classified using a hyper plane which maximizes the separation between data points and the hyper plane is constructed with the help of support vectors.

## 2. RELATED WORK

To better understand of Event Identification, it is useful to review and examine the existing research works in literature. Therefore, recent approaches and methodologies used for Event Identification have been discussed.

**Konstantinos N. Vavliakis et al [2013]**, has proposed an efficient event detection methodology for performing event detection from large time-stamped web document streams. The methodology successfully integrates named entity recognition, dynamic topic map discovery, topic clustering, and peak detection techniques. In addition, they proposed an efficient algorithm for detecting all important events from a document stream. The proposed methodology can be applied for discovering and summarizing interesting events to all web users. The incorporation of named entity and topic identification into the core of their method allows the presentation of personalized information to people that is interested only in certain entities or topics. In addition, this algorithm allows the representation of events at different levels of granularity. They were partially able to reduce the rate of false positives by using gazetteers, constructed from open linked data; however, such solution poses other difficulties. The results of the study provides: a) accurately detects important events, b) creates semantically rich representations of the detected events, c) can be adequately parameterized to correspond to different social perceptions of the event concept, and d) is suitable for online event detection on very large datasets. The expected complexity of the online facet of the proposed algorithm is linear with respect to the number of documents in the data stream.[1]

**Takeshi Sakaki et al [2010]**, focused on the investigated real-time nature of Twitter, in particular for event detection. Semantic analyses were applied to tweets to classify them into a positive and a negative class. They consider each Twitter user as a sensor, and set a problem to detect an event based on sensory observations. Location estimation methods such as Kalman filtering and particle filtering are used to estimate the locations of events. As an application, they developed an earthquake reporting system, which is a novel approach to notify people promptly of an earthquake event. Micro blogging has real-time characteristics that distinguish it from other social media such as blogs and collaborative bookmarks. This system detects earthquakes promptly and sends e-mails to registered users. Notification is delivered much faster than the announcements that are broadcast by the Japan Meteorological Agency (JMA).[2]

**Hassan Sayyadi et al [2009]**, has proposed a new event detection algorithm which creates a keyword graph and uses community detection methods analogous to those used for social network analysis to discover and describe

events. Constellations of keywords describing an event may be used to find related articles. This algorithm is also used to analyze events and track stories in social streams. In their community detection algorithm, nodes can fall into different communities as a word or phrase can be in keywords list of more than one event. In the current version of the algorithm it count all keywords in one community as keywords for the event, though a subset of keywords may be better, especially in cases where the number of nodes is large. In addition, while the keyword graph is a weighted graph, in order to find the betweenness centrality score, shortest paths are found on an un-weighted graph.[3]

**Wei Chen et al [2011]**, has proposed a naive implementation approach to extract a hot spot of a given topic in a time-stamped document set. Topics can be basic, containing a simple list of keywords, or complex. Logical relationships such as and or, and not are used to build complex topics from basic topics. A concept of presence measure of a topic based on fuzzy set theory is introduced to compute the amount of information related to the topic in the document set. Each interval in the time period of the document set is associated with a numeric value which called as discrepancy score. A high discrepancy score indicates that the documents in the time interval are more focused on the topic than those outside of the time interval. A hot spot of a given topic is defined as a time interval with the highest discrepancy score. It first describes a naive implementation for extracting hot spots. It then construct an algorithm called EHE (Efficient Hot Spot Extraction) using several efficient strategies to improve performance. It also introduces the notion of a topic DAG to facilitate an efficient computation of presence measures of complex topics.[4]

**Peter Kolb et al [2008]**, has proposed DISCO, a tool for retrieving the distributional similarity between two given words, and for retrieving the distributionally most similar words for a given word. Pre-computed word spaces are freely available for a number of languages including English, German, French and Italian, so DISCO can be used off the shelf. The tool is implemented in Java, provides a Java API, and can also be called from the command line. The performance of DISCO is evaluated by measuring the correlation with WordNet-based semantic similarities and with human relatedness judgments. The evaluations show that DISCO has a higher correlation with semantic similarities derived from WordNet than latent semantic analysis (LSA) and the web-based PMI-IR.[5]

### Overview of Existing system:

➤ In existing system, the learning algorithms are considered to identify and detect the unknown important topics from the large text document. Event in social web documents are identified through named entity recognition and topic modeling. The methodology in the

work integrates named entity recognition, dynamic topic map discovery, topic clustering, and peak detection techniques. The named entities are used to recognize the persons, organizations, locations, and other entities based on the specified topics. The topic discovery approach is used to create a model initially for all the documents. It is used to build a model for unlabeled documents also and hence, relevant terms are extracted. This is focused on extraction topics from documents and it requires a priori specification of the number of total topics. This problem is even more intense in the case of person entities, where maintaining a complete list of names is extremely difficult. Automatic selection of the optimized values of the algorithm's parameters based on user type is not performed. Dynamic selection of topics are not generated. Semantically related rich information of events is not detected. False positive rates are high. Disambiguation problem is also addressed. Hence the performance degradation is occurred. Hence, to overcome all these issues the Support Vector Machine (SVM) algorithm is utilized in the current work.

### 3. PROPOSED METHOD

In proposed scenario, supervised learning algorithm named as SVM algorithm is used to overcome above mentioned issues effectively. For dynamic selection of topics SVM algorithm is introduced to generate the events. To compare existing clustering algorithm and Ranking SVM is compared on the task of inducing an appropriate similarity measure for event identification on the basis of training data consisting of event assignments. The performance of the Ranking SVM is much more robust compared to the performance of the existing algorithm. The ranking SVMs can learn very good models with a small number of training examples. In this proposed system the documents are collected from several sources and apply SVM training to create a model. This algorithm is used to classify the events and new random or dynamic events efficiently. The ranking score is to be calculated based on the nearest neighbor and similarity concept. The similarity value is calculated as follows. Here  $d$  and  $e$  represents documents and events respectively.

$$\forall d_i \forall e' \neq e(d_i)$$

$$\text{sim}(d_i, e(d_i)) > \text{sim}(d_i, e')$$

or

$$\forall d_i \forall e' \neq e(d_i)$$

$$(\text{sim}(d_i, e(d_i)) - \text{sim}(d_i, e')) > 0$$

So the document topics are classified based on the similar topics and if new topics arrived it can be compute based on the label data. Hence it can achieve the high similarity and semantic concept for dynamic topics successfully. Based on the letter, word, sentence and topics the information with semantic meaning is extracted by using SVM algorithm. Within the threshold value the classification performance has been implemented. Hence, the proposed system can be able to provide more accurate result than existing scenario. It achieves dynamic selection of topics which has been generated efficiently. It is used to reduce the false positive rate significantly. Disambiguation issues are handled. Similarity Search is improved. As a rising subject, SVM algorithm playing an increasingly important role in the decision support activity of every walk of life.

### 4. Methodology

The modules in the current work are listed below:

- Document stream pre processing
- Entity and topic analysis
- Topic clustering
- Detection of similarity and semantic
- Performance analysis

#### 4.1. Document stream pre-processing

In order to avoid the notorious “garbage in, garbage out” predicament, pre-processing is necessary to reduce noise, clean erroneous instances, and appropriately transform data. During pre-processing stop-words are removed in order to create meaningful topic maps later. Moreover, stemming is required to reduce the inflected and derived words to their stem/root form and map related words to the same stem.

#### 4.2. Named Entity and topic analysis

##### Named entity recognition:

Named entity recognition (NER) is the process of finding measures of specified things in running text. It can be regarded as a classification task, where the goal is to detect and classify strings of text to different classes. Two are the dominant approaches followed in NER tasks; the first is a knowledge-based approach that uses explicit resources like hand-crafted rules and gazetteers, while the second is a dynamic approach, where a tagged corpus is used to train a supervised learning algorithm.

##### Topic discovery

The objective of this step is to discover sets of topics, as expressed by a stream of documents that identify their semantic content of those documents and express the

semantic similarity among them. Topic modelling is done by the use of Latent Dirichlet Allocation (LDA) for extracting semantic information from document streams.

Topic modelling is based on the assumption that each document  $d_i$  can be described as a random mixture of topics and each topic as a focused multinomial distribution over terms. LDA builds a set of thematic topics from terms that tend to co-occur in a given set of documents. The result of the process is a set of  $N_\theta$  topics, each expressed with a set of  $N_w$  terms. The number of topics  $N_\theta$  and the number of terms per topic  $N_w$  have to be defined in advance. The two parameters can be used to adjust the degree of specialization of the latent topics. LDA discovers a mixture of topics  $P(\theta|d)$  for each document  $d_i$ , where each topic is described as a mixture of terms  $P(w|\theta)$  following another probability distribution as given

$$P(W_i|d) = \sum_{j=1}^{N_\theta} P(w_i|\theta_j)P(\theta_j|d)$$

The probability of the  $w_i$  term describing a given document  $d$  is  $P(w_i|d)$ , where  $\theta_j$  is the latent topic, and  $P(w_i|\theta_j)$  is the probability of  $w_i$  within topic  $j$ . The probability of picking a term from topic  $j$  in a document is  $P(\theta_j|d)$ .

### 4.3. Topic clustering

Two popular clustering techniques that can efficiently merge semantically relevant topics are graph and hierarchical clustering.

#### ➤ Graph clustering

Topic clusters are represented as multi-graphs, where each node corresponds to a topic and each edge represents the similarity between two nodes. It has been shown that community-based semantics emerge from this graph representation and transformation process.

A Topic Model(TM) can be regarded as a hyper graph (a bipartite graph, also known as a two-mode graph) with hyper edges. The set of vertices is partitioned into two disjoint sets:

$$\text{Topics: } \theta = \{\theta_1, \theta_2, \dots, \theta_N\}$$

$$\text{Concepts (words-tags): } C = \{w_1, w_2, \dots, w_{w_n}\}$$

Thus, the topic model TM is defined as  $TM \subseteq \theta \times C$ . The bipartite network is defined as  $H(TM) = \langle V, E \rangle$ , where  $V = \theta \cup C$  is the set of vertices, and  $E = \{ \{\theta, w\} | (\theta, w) \in TM \}$  is the set of edges. Then can reduce the hyper graph into a one-mode network by folding it. This can be achieved if we

denote the matrix of the hyper graph as  $H = \{h_{ij}\}$ , where  $h_{ij} = 1$  if-f topic  $\theta_j$  is associated with concept  $w_i$ . Then define a new matrix  $S = \{s_{ij}\}$ , where  $s_{ij} = \sum_{x=1}^k h_{ix} h_{xj}$  (or in matrix notation  $S = HH'$ ), showing the associations of topics weighted by the number of common concepts. The centrality score of an edge is defined as the extent to which that edge lies along shortest paths between all pairs of nodes, and is a good measure to find the edges between two communities. Inter-community edges will always obtain a high score, since the shortest paths between nodes from different communities will pass through them. Thus, by computing the score for all edges in the graph, the edges having highest scores are removed. The topics having similar concepts are clustered.

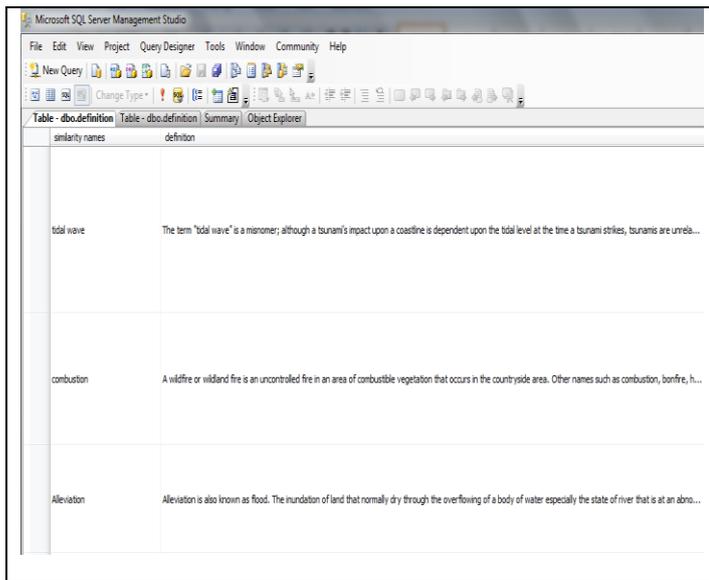
#### ➤ Hierarchical clustering

The generation of a graph comprising many topics is a resource-consuming procedure that is not suitable for real-time event detection. As an alternative to edge clustering, that can use single-link hierarchical clustering which is much faster than graph clustering, and in many cases yields similar results. Single-link is based on the nearest neighbour concept, where the similarity of two clusters is the similarity of their most similar members.

To perform single-link text clustering it is necessary to define a representation model of textual data, a similarity measure among the documents, and a strategy for the cluster formation. The widely used space-vector model represents each document  $d_i$  as a vector of words, where each term is accompanied by its frequency of occurrence. Clusters are successively merged (in each step the two clusters whose two closest members have the smallest distance) until their similarity reaches a predefined partition distance  $p$ .

### 4.4 Detection of similarity and semantic

This module, describes computing the similarity and semantic meaning for the corresponding topic query. The Word Net tool contains similarity words for particular sentence and a word also consists of synonyms. It has top  $k$  ranked sentences and uses the values accordingly for corresponding topic query. In SQL database by storing the similarity words and their meaning easily can find the particular search.



similarity names	definition
total wave	The term "total wave" is a misnomer; although a tsunami's impact upon a coastline is dependent upon the total level at the time a tsunami strikes, tsunamis are unreli...
combustion	A wildfire or wildland fire is an uncontrolled fire in an area of combustible vegetation that occurs in the countryside area. Other names such as combustion, bonfire, h...
Alleviation	Alleviation is also known as flood. The inundation of land that normally dry through the overflowing of a body of water especially the state of river that is at an abno...

Fig 4.1 Similarity Words

Word Net is a lexical database for the English language. It groups English words into sets of synonyms, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. Word Net can thus be seen as a combination of dictionary and thesaurus.

**Proposed Algorithm**

1. Consider the input query Q and R as relations
2. For i=1 → N do
3. For (inti=0; i<word length; i++);
4. WordInformation[i]=Find WordInnformation for Words(i) by Wordnet
5. For (j=1; j<R;j++)
6. R= // no.of relations exist in wordnet concept.
7. For each concept of wordnet
8. IfType(wordType.word) is a noun then
9. wordDistance= wordType.wordGetSimilarity
10. Build similarity matrix
11. IfType(wordType.word) is a noun then
12. wordDistance= wordType.wordGetSemantic
13. Improve the relevance score value based on the web page counts
14. Update all the metadata values
15. Obtain Semantic and Similarity results

**4.5. Performance analysis**

The metrics such as accuracy and time factors are used to compare the existing and proposed scenario using the performance metrics. In existing scenario, the accuracy values are lower and time complexity is high. In proposed scenario, the accuracy value is higher and time complexity is reduced significantly.

**SVM algorithm:**

1. Candidate support vector (SV) = {closest pair from opposite classes}
2. For all d
3. Top k(d) = retrieve a ranked list of promising event candidates to which d could belong
4. For all e  $\in topk$
5. Compute P(e/d) // e=event d=document
6. End for
7. While there are violating points do
8. Find a violator
9. CandidateSV = candidateSV violator
10. If any  $<0$  due to addition of c to S then
11. CandidateSV = candidateSV \ p
12. Repeat till all such points are pruned
13. End if
14. End while

This algorithm is used to improve the topics similarity and semantic concepts and prediction results are higher in accuracy. Finally concluded the proposed work yields superior performance rather than existing work.

**5. RESULTS AND DISCUSSION**

**Time factor:**

In computation, the algorithms are estimated to reduce the time complexity. For number of files the existing and proposed algorithms are executed in various time factor values. The less time execution values called higher performance in the scenario which is provided by using proposed algorithm.

No of files	NER Clustering (time in sec)	SVM Algorithm(time in sec)
Event detection		
4	5	4
5	6.5	4.5
6	7	5
7	7.5	6.5
9	10.5	7.8

**Table 5.4 Time Factor**

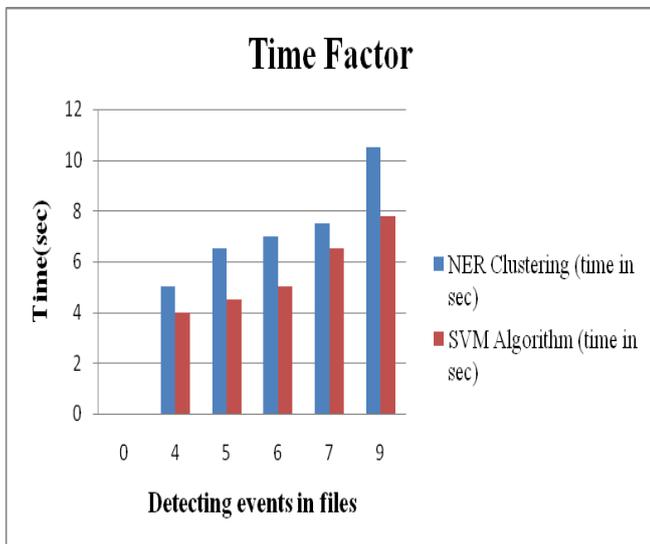
experimental result, proposed scenario is better than to existing scenario using SVM algorithm.

**FUTURE WORK**

Further, the research has also been continued by the means of online search using SVM classification algorithm. Also research can be extended in terms of huge dimensional dataset along with number of topics by using various advanced techniques in data mining to make the web search more reliable and user friendly.

**REFERENCES**

- [1] Vavliakis .K.N et al, Event identification in web social media through named entity recognition and topic modeling / Data & Knowledge Engineering 88 (2013) 1-24.
- [2] Sakaki .T, Okazaki .M, Matsuo .Y, Earthquake shakes twitter users: real-time event detection by social sensors, Proceedings of the 19th International Conference on World wide web, WWW '10, ACM, New York, NY, USA, 2010, pp. 851-860.
- [3] Sayyadi .H, Hurst .M, Maykov .A, Event detection and tracking in social streams, 3rd Int'l AAAI Conference on Weblogs and Social Media, ICWSM'09, AAAI, 2009.
- [4] Wei Chen and ParvathiChundi, Extracting hot spots of topics from time-stamped documents, Data & Knowledge Engineering 70 (2011) 642-660.
- [5] Kolb .P, DISCO: a multilingual database of distributionally similar words, in: A. Storrer, A. Geyken, A. Siebert, K.-M. Würzner (Eds.), KONVENS 2008 —Ergänzungsband:Textressourcen und lexikalischesWissen, 2008, pp. 37-44.



**Figure 5.4 Time Factor**

**5. Conclusion and Future Work**

The proposed system provides superior performance rather than existing scenario is concluded. In existing system, the methods are named as LDA clustering and Named Entity Recognition (NER) is introduced. These methods are used to retrieve the important topics and similarity values for corresponding topic query. However it has time complexity issues and accuracy values are also not achieved significantly. To overcome this in the proposed scenario, supervised learning algorithm such as Support Vector Machine (SVM) is introduced. This algorithm is used to classify the higher level similarity and more relevant topics with several meaningful results. It increases the accuracy values using Word Net concept and reduced the time complexity values. From the