Scientific Journal of Impact Factor (SJIF): 4.14

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 11, November -2016

PREEMINENT FEATURE DISCOVERY AND DOCUMENT CLUSTERING USING TEXT MINING

[1]S.Nithya, [2]Mr. N.Kamalraj

[1]M.Phil Scholar, Dept. of Computer Science, Dr.SNS College of Arts and Science, Coimbatore, Tamil Nadu, India
[2]Assistant Professor, Dept. of Information Technology, Dr.SNS College of Arts and Science, Coimbatore, Tamil Nadu, India

ABSTRACT- Text document classification and indexing is the major part of document management, where every document should be identified by its key terms and domain knowledge. Based on the domain knowledge, the documents are classified into different classes. For document classification there are several approaches were proposed in existing system. But the existing system is either term based or pattern based. And those systems suffered from polysemy and synonymy problems.

To make a revolution in this challenging issue, the proposed system presents an innovative model for relevance feature discovery and document classification. It discovers both positive and negative patterns in text documents as higher level features and deploys them over low-level features (terms). It also detects the most appropriate features based on its weight and semantic nature and performs the document classification. Using this approach, the document index terms, patterns and category can be identified easily. In order to perform the above, a hybrid approach is used which contains the following algorithms. (a). **sequential semantic pattern mining algorithm** for sequential pattern extraction (b). **Semantic Weighted feature ranking algorithm** to rank the higher supported terms in the form of semantic and patterns. (c). a **Rule based ontology** concept which helps to classify the documents under various classes and helps to detect the possible index terms. This helps to reduce the training data collection problems.

KEYWORDS: Text Mining, Document Classification, Clustering , Pattern .

1. INTRODUCTION

Document Classification (DC) is the process of analyzing a set of documents and labeling each one of them with an appropriate category according to its relevance towards one of a pre-defined set of categories. The field of DC is rife with potential for several modern services document centric applications, such as Document Summarization, essay scoring, organizing documents for query based information dissemination, email management and topic specific search engines.

The phenomenal rise of web based services such social networks; e-commerce and a variety of e-portals such as those for e-governance require well-organized management of documents. The first step for achieving this is an accurate, reliable and fast classification of relevant documents into a set of known categories.Emerging applications such as text based event detection, sentiment analysis and disaster management demand that Document Classification take into cognizance the meaning or semantics that are knitted into such small messages. These new paradigms pose the challenge of analyzing tightly worded documents with highly meaningful content. This thesis focuses on the problems of *Text Based Document Classification and Analysis*. The main aim is to investigate the urgent approaches and challenges that exist in the field of DC and enhance its power by developing a context based approaches that cork in symphony with traditional statistical approaches.

1.1 SCOPE OF THE RESEARCH

As document clustering and classification plays a very vital role in various applications, many researches are still being done. The upcoming innovations are mainly due to the properties and the Attribute of existing methods. These existing approaches form the basis for the various innovations in the field of document management. From the existing clustering and classification techniques, it is clearly detected that the clustering techniques based on GA, fuzzy and ontology provide significant results and performance and SVM from classification. Hence, this research concentrates mainly on the semantic enrichments with ontology concepts for better performance. The thesis aims to fulfill the objectives of various cluster related algorithms by developing an effective and efficient document semi supervised classification technique. That technique is expected to give good accuracy and performance.

From the existing clustering, feature extraction techniques; it is clearly observed that the data mining techniques based on EM, pattern mining and ontology provide significant results and performance. Hence, this proposed system concentrates mainly on the sequential semantic pattern discovery, which eliminates the existing pattern discovery issues.

The proposal is based ontology clustering for better performance. This thesis aims to fulfill the objectives of various cluster and pattern discovery related algorithms by developing an effective and efficient document management technique. That technique is expected to give good accuracy and performance.

1.2 METHODOLOGY

A novel approach becomes necessary in the areas of document classification which provides higher efficiency and accuracy. The research study deals primarily with five proposed approaches for document classification and management. They are:

- Effective use of both relevant and irrelevant feedback (positive and negative feedback) to find useful features;
- Integration of both term and pattern features together rather than using them in two Partition stages.
- So proposed system is hybrid architecture.
- Proposes a **sequential semantic pattern mining algorithm** for feature extraction
- **Semantic Weighted feature ranking algorithm And Rule based ontology**

Each of the proposed document clustering methods is evaluated on the following criteria-Classification accuracy, Objective function, Classification time, and Algorithm complexity, Speed up, Entropy and Separation index.
The improvements achieved in those performance measures have been tested for statistical significance using different functionalities.

2. LITERATURE REVIEW

2.1 Statistical Approaches

Statistical approaches are used to classify documents using frequencies of tokens. In this section, some very popular statistical approaches for DC are described.

2.1.1 Naive Bayes Classifier

A Naïve Bayes classifier appeal Bayesian statistics with strong independence assumptions on the features that drive the classification process. Essentially, the presence or absence of a unique feature of a class is assumed to be unrelated to the presence or absence of any other feature. *Bayesian spam filtering* is a form of e-mail filtering that uses the Naïve Bayesian classifier to identify spam e-mail [***11***].

The main strength of the Naïve Bayes algorithm lies in its simplicity. Since the variables are mutually independent, only the variances of individual class variables need to be determined rather than control the entire set of covariances. This makes Naïve Bayes one of the most popular models for email filtering. It is robust, continuously improving its accuracy while adapting to each user's liking when he/she identifies incorrect classifications thus allowing continuous rectified training of the model. In [***12]***, the authors constructed a corpus *Ling-Spam* with 2411 non spam and 481 spam messages and used a parameter l to induce greater penalty to false positives. They demonstrated that the weighed accuracy of a naïve-Bayesian email filter can pass 99%. Variations of the basic algorithm for example, using word positions and multi-word N-grams as attributes have also yielded good results [***13***]. However, the Naïve Bayes classifier is susceptible to *Bayesian poisoning*, a state where a spammer mixes a large amount of legitimate text or video data to get around the filter's probabilistic detection mechanism.

2.1.2 Support Vector Machine (SVM)

An SVM is a supervised learning method based on *structural risk minimization* [14]. It subjects every category to a separate binary classifier. SVM's forte is that it is relatively immune to the depth of the feature space, focusing instead on *maximizing the margin* between positive and negative examples of training documents. It avoids the use of many training documents; appoint only those near the classification border, to construct an irregular border separating positive and negative examples. By employing a suitable kernel function, it can learn polynomial classifiers, radial basis functions and three-layered sigmoid neural nets, thus acquiring universal learning abilities.

User can observe from the table that SVM reported 8% improved classification accuracy by using term frequency and HTML tags over only Term Frequencies used as features.

Hong and Cho employed semi-supervised SVM for classification of unlabeled documents [14]. They initiated their experiments on a small set of labeled documents. SVM was trained by this small set initially and then used to categorize unlabeled documents. Since, it could not perform on whole dataset at once, sampling was required. SVM classified these samples and then kept them in a labeled document set. The process was repeated until it satisfied the termination condition. They performed their experiments on Reuter corpus.

2.1.3 K-Nearest Neighbors (KNN)

The KNN technique [***15***] proceeds by choosing first random data points as initial *seed* clusters. Next, it enters a learning phase when training data points are iteratively assigned to a cluster whose center is located at the nearest distance (*e.g.* Euclidean distance). Cluster centers are repeatedly adjusted to the mean of their currently acquired data points. The classification algorithm tries to find the K nearest neighbor of a test data point and uses a majority vote to determine its class label. The performance of KNN classifier is primarily determined by (i) an appropriate choice of *K* which can be quite tricky if either the data is non-uniformly distributed or if there is noisy data, and (ii) the distance metric applied. The value of *K* may need to be tuned for a given application.

In [*17*], Nako*vet al*applied latent semantic analysis and KNN classification with 10 fold cross validation to two document collections: *Ling Spam* corpus [***18]*** and a personal collection of emails containing 940 non spam and 525 spam messages. They achieved an accuracy of 99.65% for moderate values of *K* set to 3 and 4. In [***19***], the authors applied KNN to the SA2 corpus with 10 point validation. They demonstrated that when the value of *K* is set at 3, the overall accuracy is 93% with a distinct split in accuracy at 98.6% for good email and 79.8% of spam mail.

In [***18***] Zahedi and Sarkardei have modified the *TF_IDF* for feature weighting with *mutual information (MI)* to classify text to a set of categories. *MI* considers the distribution of the features in different categories while weighting each feature in each category. They used a modified KNN algorithm as a classifier. According to the authors, if the frequency of term *Ti* of a document *Di* is distributed over other documents belonging to same category *Ci*, the weight of *Ti* should be increased in *Ci*. Whereas, if the number of categories containing the term *Ti* increases except category *Ci*, the weight of *Ti* should be decreased in category *Ci*. After this feature weighting step, the authors applied modified-KNN on these features and reported 91% average accuracy and 89% average recall.

2.1.4 Fuzzy Logic

Fuzzy logic uses linguistic variables, overlapping classes and approximate reasoning to model a classification problem [***20***]. The works in [***21***] [***22***] show that fuzzy logic lends well to spam detection as indeed the classes *spam* and *non-spam*messages overlap over a fuzzy boundary. Sayed*et al* employ fuzzy-based spam detection by first pre-processing the documents building a fuzzy-model of overlapping categories **{spam, valid}** with membership functions construct from the training set and**,** and classifying input messages by calculating the fuzzy similarity measure between the received message and each category [***23***]. The authors tested their classifier with various fuzzy conjunction and disjunction operators using 4 datasets, two for training and two for testing. Averaging over the 4 cases, the best results were obtained for Bounded Diff. with an accuracy of 97.2%, spam recall of 90.5% and spam precision of 97.6%.

In paper [***24***], Kim *et al* retained hyperlinks because spammers can minimize text but list hyperlinks. They demonstrate that feature selection by fuzzy inference is superior to conventional methods such as Information Gain. This indicates that the linguistic modeling in fuzzy logic is well-suited for both feature extraction and DC.

2.2) Collaborative Approaches

Term weighting has applications in question answering and information extraction. In existing, the authors adopted a combined approach by integrating both statistical and NLP based features to define a term weighting *context function* that progressively refines the weights of terms in a document based on the influence of surrounding words. Each term weight or score is recursively evaluated by a combination of its current score denoting its implicit relevance and by the context function that generates the influencing score.

In [***29***], the authors extract groups of four to six words (*N*-grams) that repeat at least twice. These multi-words form tokens that define the composite features owning to multiple instances (statistical indicator) of their co-occurrence (contextual indicator). They reported .8156 avg-precisions, .8215 avg-recall and .8156 avg-F-measure on Reuters-21578.

3. RESEARCH METHODLOGY

3.1 Contributions

The proposed Document Classification (DC) system is called "SRO (Semantic Rule based Ontology for DC)" because it uses two sources of existing knowledge, and ontology determine the semantic content of categories and documents and map them. **Figure 3.1** depicts the overall architecture of the SRO based DC scheme proposed in this chapter. The main contributions of the scheme are highlighted below.

A novel approach becomes necessary in the areas of document classification which provides higher efficiency and accuracy. The research study deals primarily with five proposed approaches for document classification and management. They are:

- Effective use of both relevant and irrelevant feedback (positive and negative feedback) to find useful features;
- Integration of both term and pattern features together rather than using them in two partition stages.
- So proposed system is hybrid architecture, which is named as SRO.
- SRO contains the sequential semantic pattern mining algorithm for fast and reliable feature extraction
- Semantic Weighted feature ranking algorithm for keyword ranking for effective term detection in DC
- Rule based ontology detects the domain and sub domains of the document and finds the similarity score.

The improvements achieved in those performance measures have been tested for statistical significance using different functionalities.

Fig 3.1 document classification process

The chapter initiates the process of classification by populating each category with similar concept terms that define the feature space of the category. Semantic process expresses the meaning of each sense of a base word with its Synonym set. For each non-trivial token which is present in a document and for each category name, their lexically related terms can be found in word dll starting with their relevant senses and transitively following different relation types. Lexical cohesion refers to a range of textual cohesion that allows the use of similar meaning words on Synonyms, generalization the concept on Hypernyms, specialized versions of a concept called Hyponyms or constituent parts of an object called Meronyms. Terms that share a common Hypernym are called Coordinate terms. The set of lexically related terms of a base word are called Lexical Semantics. The concern about generating a suitable list of keywords for a given category can be addressed by extracting all Lexical Semantics of the category from the WordNet.

Positive Term Detection: the next concern is to extract meaningful words from a document. For accomplishing this, the SRO uses a feature discovery (FD) process, which can find the both positive and negative term and term sets from the set of documents. The fundamental premise of *FD* is that if two tokens in a document have a common list of Lexical Semantics, then the corresponding features represent the document's meaningful intent greatly. Therefore, such tokens can be retained and other tokens can be pruned as their presence is incidental. This reduces the feature space of each document.

Topic Detection: The tokens of the reduced documents are matched with keywords of each category. Based on the Keyword Strengths and term frequencies of matched tokens, the system finds the topic of the given document d. The topic belongs to a document to each category is calculated. Finally, the document is ascribed to the category with highest *similarity* value.

SRO aided Keyword Enrichment: With the help of the Reuters dataset, the algorithm enriches category keyword lists by using the tokens of classified documents. Reuters is a free, open content online repository created through the collaborative effort of a community of users**.** Reuter's dataset are prepared with n number of topics. It has functions to handle many fundamental tasks in computational linguistics. This hit these powerful features of UCI to argument more keywords and enhance the feature space of each category. The further chapter gives more detail on the proposal work in subsequent sections.

3.2 Framework for SRO

This Section organizes the methodologies and algorithms involved in the proposed system.This Work focuses on the problem of automatically extracting and classifying text documents based on the weighted positive and negative key terms and based on the sequence, it can detect the exact similarity between the documents. The stepwise detailed description of the SRO system now follows. The fig 3.2.1 shows the overall process of the proposed work.

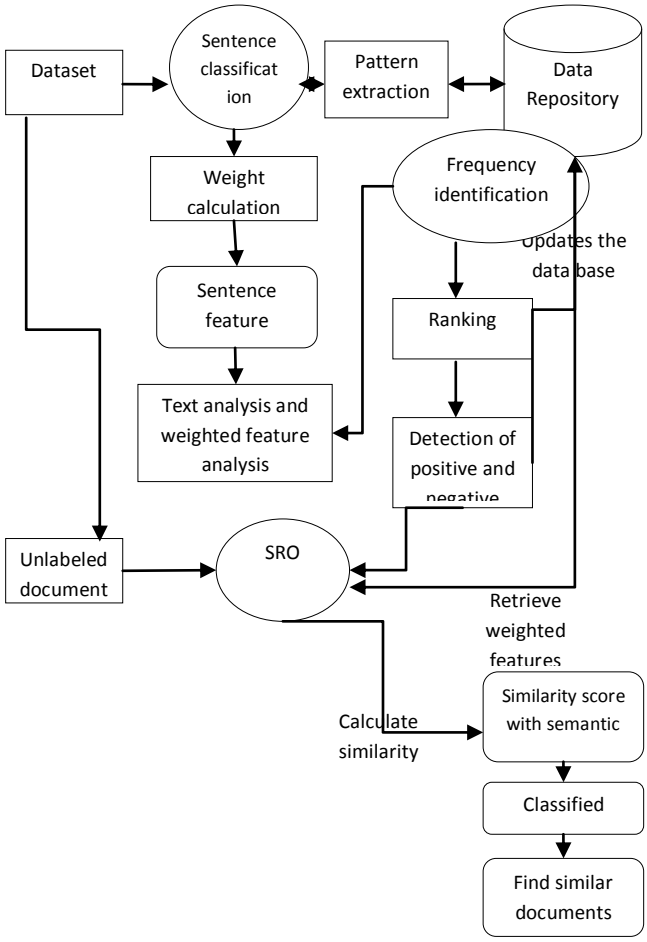

Fig 3.2.1 SRO architecture in document classification

Feature discovery algorithm
Algorithm: FD
Input: 1. A set of tokens in a pre-processed document D**.**
2. Feature Threshold T_{FD}.
Output: A set of semantically related terms in new document SD

Steps:

1. Start the FD process and declare SD=null
   a. SD=0
2. **For** each token($\forall Token\ T$)
   a. $w \in D$
3. Obtain Lexical Semantic Set **Sw**fromWord dll;
4. **For** each token $w\in$ D**{**
5. **For** each other token z^1w, zÎ D
   a. **If** (Sw∩Sz≥P_{FD}) **then**
      i. SD=SD ∪{w, z}
   b. End if
6. Delete D
   a. Return SD
7. **End**

By FD, the algorithm extracts strongly related words of the document, which are important for classification, which is also known as positive term. It inputs each token w of a document to the Worddll to obtain its Lexical Semantics viz., Synonyms and Coordinate terms. This is termed as its Lexical-semantics set Sw. Now, starting with first token, the system takes each token w and finds the intersection of Sw with those of all other tokens in the document. If the intersection exceeds predefined threshold P_{FD}, the token is retained. Otherwise it is dropped from the document. The tokens thus collected from the pre-processed document signify the document's meaning to a greater extent and reduce the dimensionality of document representation. Let Nwbe the final number of tokens representing the document.

3.3.1Algorithm: Ontology Generation
Input: Starting concept CS of concept lattice F (K) and a similarity threshold TS
Output: A set of generated conceptual groups SC
Steps:
1: SC {}
2: F(K) An empty concept lattice
3: Add CS to (K)
4: for each sub-concept of CS in F(K) do
5: (C)Conceptual_Group_Generation(C, (K), TS)
6: if E(CS,C) = < TS then
7: SC SC{ (C)}
8: else
9: Insert (C) to (K) with sup (F (K)) as a sub- concept of CS
10: endif
11: endfor
12: SC ←SC{ F(K)}

A formal concept should belong to one minimum group generated through concepts, but it can also be more than a single concept group. This attribute is obtained from the characteristic of concepts that an object can belong to more than a single concept.

4. IMPLEMENTATION AND RESULTS

4.1 EXPERIMENT SET UP

a. **Software and Platform:** The SRO framework described above was coded using C#.Net version 4.0 for initialization, keyword database pruning and feature vector preparation tasks. The C#.Net word libraries are used for semantic based evolution. The .Net framework and its library provide numerous advantages, the proposed system is a dynamic one, and datasets are not static. User can use own documents for the evaluation. MSSQL Connecter was established to make database queries to SRO. The software was run on an i7 quad core processor 2.4GHz with Windows 7.

Data Sets: The proposed system used real-time and synthetic datasets. Different corpus adopts different rules and models. Some have documents with specialized vocabulary containing words that are repeated frequently. On the other hand, corpus derived from certain sources exhibit creative writing style with word occurrences seldom repeated in their documents. The objective was to achieve a corpus specific combination of statistical and context features that gives the most accurate classification for varying writing styles and the average size of documents. In order to validate the efficacy of the approach on varying corpora, the proposed system experimented different sources for the experiments.

1) **Reuters 21578 dataset:** Currently the most widely used test collection for text categorization research, though likely to be superseded over the next few years by RCV1. The data was originally collected and labeled by Carnegie Group, Inc. and Reuters, Ltd. in the course of developing the CONSTRUE text grouping system. Further details, including discussion of previous versions of the collection (e.g. Reuters-22173), are available in the website. The dataset is available at http://www.research.att.com/~lewis/reuters21578.html,ftp:://canberra.cs.umass.edu/pub/reuters.It has 90 specialized categories. All the 90 categories can be used in the experiments.

2) ***Synthetic Domain related Datasets:*** in the proposed system, the synthetic dataset are also used. This handcrafted dataset which is known as *Domain corpus* containing selected research domains, articles from the IEEE explore. This corpus has several categories: *Knowledge and data engineering, Image processing, networking and security etc*. Before applying the SRO method to a large data set, there is a need to assess its performance on a prototype

corpus with smaller number of documents. So, therefore generated two data sets for each of the above corpora by extracting documents randomly and assigning them to the following datasets:

a) Dataset1: For each of the afore-mentioned sources, the prepared a Dataset1 comprising 12 documents with 3 documents in each of the four respective categories. Among the 3 documents in each category, 2 of them were selected for training the classifier and the remaining 1 document were used for testing the classifier.

b) Dataset2 For each of the three corpuses, the Large Datasets comprised 120 documents with 40 documents in each of its four categories. Among the 40 documents in each category, 70 documents were selected for training the classifier and the remaining 50 were set aside to test the classifier.

Table 4.1 dataset summary

Data sets	Average document size (No of tokens)	Number of documents	Average TF_IDF per document	Average Context score per document
R21578	657	120	0.12	0.0748
Dataset1	123	12	0.07	0.1161
Dataset2	808	120	0.09	0.1622

In this chapter, the SRO implementation and performed an analysis of each dataset to make a comparative assessment of their degree of repetitiveness and contextual content. **Table 4.1** summarizes the datasets. Column 1 shows the names of the datasets. Column 2 indicates the average size of the tokenized documents. Column 3 gives the total number of documents in the experiment. Column 4 gives an overall measure of the statistical content in the corpus in terms of average TF_IDF per document denoted as t. Column 5 gives a measure of the overall contextual content of the corpus in terms of average context score per document in the corpus, denoted as k. this will use these values to assess how appropriately the SRO assigned the semantic value in each case.

4.2 Exploration on different Datasets

The next process depicts results of applying the SRO on each of the chosen Datasets. In each case, this plots the classification accuracy, along consecutive generations of the effective feature driven evolution process promulgated by the SRO system. For ease of reference and for comparison, this has collects the optimal weights that were experimentally obtained for all statistical, Lexical Semantic and features at the end of the exploration process for sample dataset is given in **Table 4.2.**

Dataset	document clustering is the process of grouping of text mining is an iterative process, which gives high information and effective data in the text mining domain using pattern analysis and pattern clustering
Total sentences	2
Total Terms	32
Total Characters	211
TF-IDF	an [1] analysis [1] and [2] clustering [2] data [1] document [1] domain [1] effective [1] gives [1] grouping [1] high [1] in [1] information [1] is [2] iterative [1] mining [2] of [2] pattern [2] process [1] process, [1] text [1] the [2] using [1] which [1]
Stop word elimination	document clustering process grouping mining iterative process which gives high information effective data text mining domain using pattern analysis pattern clustering
TF-IDF after stop word elimination	analysis [1] clustering [2] data [1] document [1] domain [1] effective [1] gives [1] grouping [1] high [1] information [1] iterative [1] mining [2] pattern [2] process [2] text [1] using [1] which [1]
Sequence Detection	document Start: 0 End: 8 clustering Start: 9 End: 19
Weighted Keywords	Clustering, Mining, Patterns
Semantic Keywords	Category, group, extract

Table 4.2 results obtained at every stage for sample dataset

The performance of the proposed algorithm is measured based on the detection time, accuracy, similarity and unlabeled data handling. The followings are the initial performance analysis of the above sample dataset results.

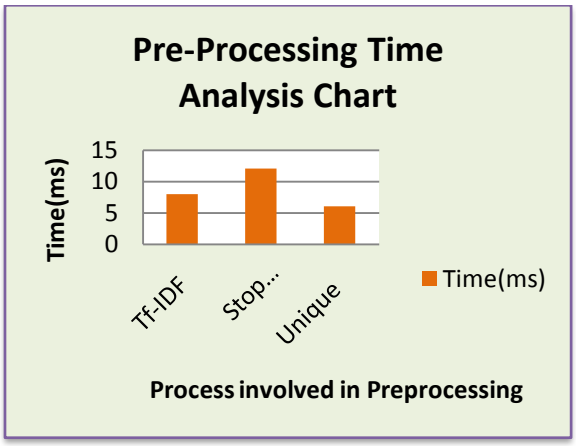

Fig 4.1 Pre-Processing time analysis chart

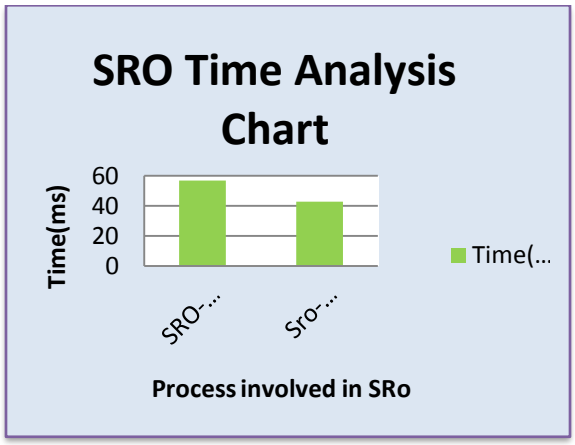

Fig 4.2 SRO time analysis chart

For the above given sample dataset, the time is evaluated. The fig 4.1 shows the preprocessing time in the SRO process. That includes the term detection, frequency and stop word elimination processes. The fig 4.2 shows the time taken analysis chart for the SRO training and test process. The proposed semi supervised document classifier takes less time for test process and slightly high for training process. The test process includes the weighted semantic feature matching, so the results are more accurate and fast.

4.3 Performance comparison

Assessment of overall performance

In this subsection, the report gives the results and overall performance of the SRO model. So, the first process is comparing its accuracy with that obtained by RFDT classification on the same corpora. Then the next illustrate the variety of prior solutions present in the final iteration. Finally this gives the salient performance parameters of the best feature obtained for each dataset and compare them with previously reported results.

Comparison with RFDT Model

In order to compare with the SRO approach with a RFDT, this chapter conducted RFDT based document grouping process using only *TF_IDF* features of each document in each of the data sets. The existing RFDT was trained and tested for each corpus. **Table 4.3** tabulates the accuracy results obtained for the two approaches.

a) The proposed collaborative approach performs comparatively better than RFDT for both datasets of each corpus. The average accuracy for the collaborative method is 96.5% as compared with 84.3% with the RFDT method, thus giving an improvement of 25%.

Table 4.3 Performance Comparison between RFDT and SRO approaches

Datasets	Accuracy using RFDT (%)	Accuracy using SRO (%)
R21578	86	96.5
Dataset1	84.5	97
Dataset2	83	96

Table 4.3

b) In cases where the RFDT method gave acceptable results, *i.e.* 86% for the R21578 dataset and 84.5 % for the Dataset1, the SRO approach enhanced it in both cases to 96.5% and 97% respectively.

c) For the synthetic and large dataset 2, the RFDT approach led to rather poor results which were dramatically improved with a collaborative approach. For instance, the classification accuracy of the Dataset2 was only 83% using a RFDT approach. This improved to as much as 96% with the SRO approach. This is because the DC system was able to utilize the context based featured maximally in the domain corpus.

4.4 Recall and Precision Analysis

Figure 4.3 exhibits the proposed systems accuracy, which presents in the final iteration produced by the SRO for the first dataset, in terms of two conflicting objectives precision and recall. Solution 1 has the highest precision at 87.77% but poorest recall at 78%. Solution 5 on the other hand has the best recall at 87.77% but least precision at 78%. Solutions 2, 3 and 4 are positioned in between. The SRO does give the benefit of tradeoff choices between precision and recall. The user has the flexibility to choose a solution that best suits the target application.

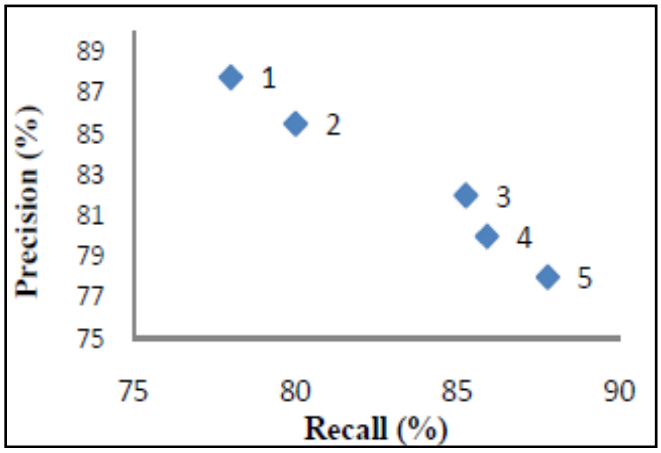

Figure 4.3 precision and recall for the SRO process

Accuracy, Precision, Recall & F1-Measure

Table 4.4 shows the overall performance of the proposed scheme. It shows the micro-averaged accuracy, precision, recall and F-measure on the above specified 3 ifferent datasets. The RFDT ranking processin base paper shows an average precision of 81.56%, an average recall of 82.15% and an average F-measure of 81.56% on Reuters 21578 datasets. In comparison, as shown in Table 6, the SRO gave an average precision of 95.24%, an average recall of 94.75% and an average F-measure 95.41%.

Datasets	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
R21578	96	88.56	96	92
Dataset1	100	100	100	100
Dataset2	97.7	100	97.3	98.9

Table 4.4

In above table shows the outcome of the proposed SRO model and its various parameters are discussed in the table 4.4. The F-measure of the proposed system is approximately 92% (as observed in the graph) on R21578 dataset and approximately 98.9% on a small dataset using their large dataset Dataset2. The main process of the SRO model it reduces the unwanted features, In contrast with the existing works that rely only upon multi-word features within a document, SRO have utilized the highly structured network of words connected by their lexical relationships as stored in the Ontology and the well-organized categorical information compiled in semantic dictionary. The above results reveal that a SRO approach that garners support of ontology databases such as WordNet and Semantic to perform contextual text analysis and leverages it with RFDT have a clear edge over past attempts at combining the two approaches by interpreting contexts in the form of word-groupings present in the same documents.

Accuracy calculation:

The system finally performs the analysis to show the accuracy of the proposed system. Accuracy refers to the Matching of data classified an accurate type in total data, namely the situation TP and TN, thus the accuracy is

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} * 100\%$$

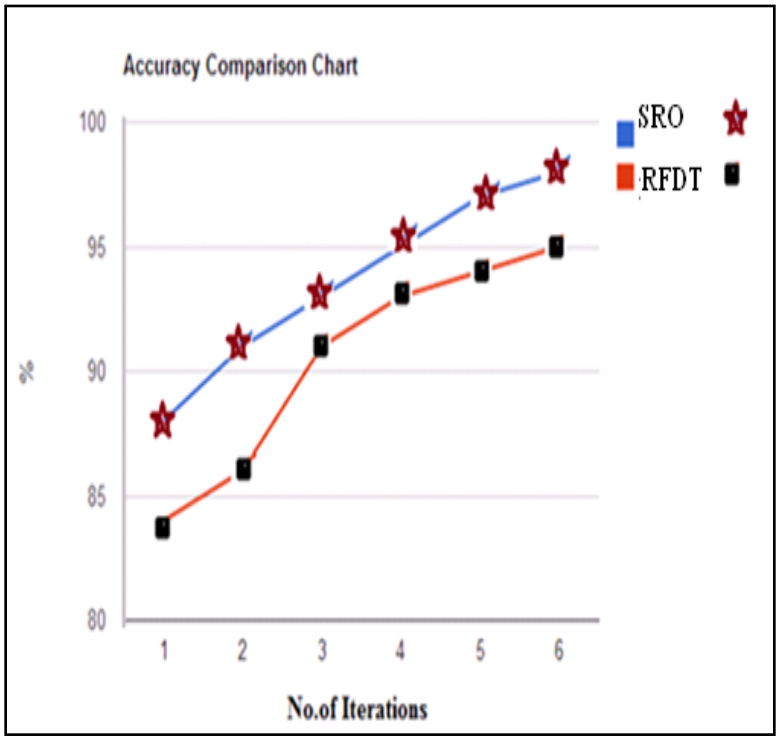

Fig: 4.4 Performance and accuracy analysis

Detection and identification of high ranked features and most negative features and its frequency can be generalized as the following Table 4.5:

- True positive (TP): the count of documents detected correctly.
- True negative (TN): the count of documents detected when it is actually not a specific domain.
- False positive (FP): The count of documents detected as irrelevant when it is actually relevant one, namely false alarm.
- False negative (FN): The count of documents detected as relevant one when it is actually irrelevant, namely the documents which can be detected by SRO system.

Nowadays, document classification in system requires high detection rate and low false alarm rate, thus the research compares accuracy, detection rate and false alarm rate, and lists the comparison results of various documents.

Table: 4.5. Performance comparison table

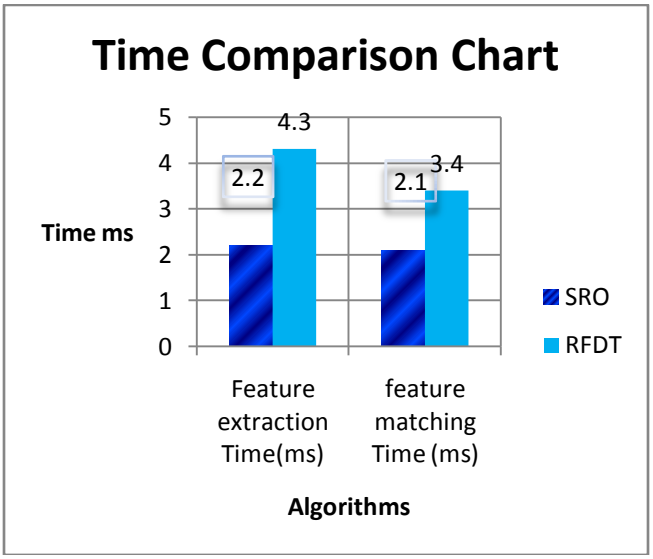

Fig: 4.5 Time comparison between existing RFDT and proposed SRO

The above figure represents the comparison between existing and proposed system based on the Training time. This chapter compared the training time of the other classification algorithms with the SRO.

5. CONCLUSION AND FUTURE WORK

In this research work a new document classification approach is implemented. A new semantic rule based ontology (SRO) is proposed to classify the documents effectively. The SRO reduces the data collection problem by applying semantic semi-supervised learning process. The SRO is also finds effective positive and negative features for fast data grouping. Using the SRO, the two documents can be compared and the similarity can be calculated. The similarity is calculated based on both external and deep features of documents. In the existing system document classification and grouping suffers from numerous issues, such as accuracy, training overhead, feature matching time etc., the SRO is created with the aim of reducing those challenges and issues in the document classification process. In the recent scenario thousands and lacks of documents are handled by every user. So managing and retrieving those documents are very tedious. The SRO can perform document classification based on its similarity, positive and negative features and different other factors. So this simplifies the document management process. The SRO populates the high featured terms for training process, and this reduces the test time of the system. The new and enhanced semantic ontology framework increases the classification accuracy.

FUTURE WORK

Several open problems emerge from the dissertation and can serve as objectives for future research. The first idea to extend the work to the next part is applying this in various classification applications;the methods that have been developed can be adapted for specific applications. The SRO based DC scheme can be extended to group the personalized e-mail of different users according to their individual writing styles.

Metrics	RFDT	SRO
Feature extraction Time(ms)	4.3	2.2
Efficiency	Ordinary	Better
Precision (%)	90.7	97.5
Feature matching Time (ms)	3.4	2.1

The concept of *Belongingness* can be utilized for real time online context based e-mail classification. Different web portals house documents in differentiated forms such as Tweets, blogs, comments, FAQs, posts, tagged images etc. Each has its unique semantic characteristics that can be subjected to knowledge based contextual analysis for applications such as localized sentiment prediction, articles categorization etc. the next idea is enabling the SRO to integrate various contextual features.

References

[1] Freitag, Dayne. "Machine learning for information extraction in informal domains." *Machine learning* 39.2-3 (2000): 169-202.

[2] Kotsiantis, Sotiris B., Ioannis D. Zaharakis, and Panayiotis E. Pintelas. "Machine learning: a review of classification and combining techniques."*Artificial Intelligence Review* 26.3 (2006): 159-190.

[3] Pandey, Upasana, and S. Chakraverty. "A review of text classification approaches for e-mail management." *International Journal of Engineering and Technology* 3.2 (2011): 137.

[4] André F. T. Martins Dipanjan Das, "A Survey on Automatic Text Summarization,"in Literature Survey for the Language and Statistics II course , Carnegie Mellon University, November 2007, pp. 1-31.

[5] Upasana Pandey, S. Chakraverty, BhawnaJuneja, Ashima Arora, Pratishtha Jain, " Semantic Document Classification using Lexical Chaining and Fuzzy Approach", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-5, Pages 367-371, November 2011

[6] Upasana Pandey, S. Chakraverty, RichaMihani, Richa Sharma, Ruchika Arya, SonaliRathee, "Semantic Based Fuzzy Approach for Document Classification Using WordNet and Wikipedia", International Journal of 183 Computer Sciences and Management Studies , Vol.3, No.2, pp. 71-77, Dec.2011.

[7] Na Luoet a, "Using CoTraining and Semantic Feature Extraction for Positive and Unlabeled Text Classification," in International Seminar on Future InformationTechnology and Management Engineering, 2008

[8] Spink, Amanda. "A user-centered approach to evaluating human interaction with web search engines: an exploratory study." Information processing & management 38.3 (2002): 401-426.

[9] Rudy Prabowo and Mike Thelwall, "Sentiment Analysis: A Combined Approach,"Intenational Journal of Informetrics, Vol. 3, Nr. 2 (2009) , p. 143-157., vol. 3, no.2, pp. 143-157, 2009.

[10] Glynn, Dylan, and Justyna A. Robinson, eds. Corpus methods for semantics: Quantitative studies in polysemy and synonymy. Vol. 43. John Benjamins Publishing Company, 2014.

[11] S. Youn and D. Mcleod, "Efficient Spam Email Filtering using Adaptive Ontology," in International conference on Information Technology, 2007, pp. 249-254.

[12] Saruladha, K., and L. Sasireka. "Survey of text classification algorithms for Spam Filtering." Int. Journal of Innovative Trends in Engg (2012): 233-237.

[13] Miao, Yingbo, VladoKešelj, and EvangelosMilios. "Document clustering using character N-grams: a comparative evaluation with term-based and word-based clustering." Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005.

[14] Shawe-Taylor, John, et al. "Structural risk minimization over data-dependent hierarchies." IEEE transactions on Information Theory 44.5 (1998): 1926-1940.

[15] Zhang, Min-Ling, and Zhi-Hua Zhou. "A k-nearest neighbor based algorithm for multi-label classification." 2005 IEEE international conference on granular computing. Vol. 2. IEEE, 2005.

[16] Deng, Zhi-Hong, Kun-Hu Luo, and Hong-Liang Yu. "A study of supervised term weighting scheme for sentiment analysis." Expert Systems with Applications. 41.7 (2014): 3506-3513

[17] Girju, Roxana, et al. "Semeval-2007 task 04: Classification of semantic relations between nominals." Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics, 2007

[18] Webb, Steve, James Caverlee, and CaltonPu. "Introducing the Webb Spam Corpus: Using Email Spam to Identify Web Spam Automatically." CEAS. 2006.

[19] Pandey, Upasana, and ShampaChakravarty. "A survey on text classification techniques for e-mail filtering." Machine Learning and Computing (ICMLC), 2010 Second International Conference on. IEEE, 2010.

[20] Cordón, Oscar, María José del Jesus, and Francisco Herrera. "A proposal on reasoning methods in fuzzy rule-based classification systems." International Journal of Approximate Reasoning 20.1 (1999): 21-45.

[21] Lin, Wei. "Fuzzy logic voting method and system for classifying e-mail using inputs from multiple spam classifiers." U.S. Patent No. 7,051,077. 23 May 2006.

[22] Tarbotton, Lee Codel Lawson, Daniel Joseph Wolff, and Nicholas Paul Kelly. "Detecting unwanted properties in received email messages." U.S. Patent No. 6,757,830. 29 Jun. 2004.

[23] Lee-Kwang, Hyung, Yoon-Seon Song, and Keon-Myung Lee. "Similarity measure between fuzzy sets and between elements." Fuzzy Sets and Systems 62.3 (1994): 291-293.

[24] Casillas, Jorge, et al. "Genetic feature selection in a fuzzy rule-based classification system learning process for high-dimensional problems." Information Sciences 136.1 (2001): 135-157.

[25] Varelas, Giannis, et al. "Semantic similarity methods in wordNet and their application to information retrieval on the web." Proceedings of the 7th annual ACM international workshop on Web information and data management. ACM, 2005.

[26] Leacock, Claudia, George A. Miller, and Martin Chodorow. "Using corpus statistics and WordNet relations for sense identification." Computational Linguistics 24.1 (1998): 147-165.

[27] Rabiega-Wiśniewska, Joanna, et al. "Semantic relations among nouns in Polish WordNet grounded in lexicographic and semantic tradition." ÉtudesCognitives/StudiaKognitywne 11 (2011): 161-182.

[28] Barak, Libby, Ido Dagan, and EyalShnarch. "Text categorization from category name via lexical reference." Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. Association for Computational Linguistics, 2009.

[29] Kešelj, Vlado, et al. "N-gram-based author profiles for authorship attribution." Proceedings of the conference pacific association for computational linguistics, PACLING. Vol. 3. 2003.

[30] Bilenko, Mikhail, and Raymond J. Mooney. "Adaptive duplicate detection using learnable string similarity measures." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.

