Semantic similarity based web document classification using Artificial Bee Colony (ABC) algorithm

C.KAVITHA Dr.G.SUDHA SADASIVAM S.KIRUTHIKA
Department of Computer Science and Engineering
PSG College of technology
Peelamedu, Coimbatore, Tamil Nadu
INDIA
mail2kavithak@yahoo.com, sudhasadhasivam@yahoo.com, kiruthika.2728@gmail.com

Abstract:- Due to the exponential growth of information on the Internet and the emergent need to organize them, automated categorization of documents into predefined labels has received an ever-increased attention in the recent years for efficient information retrieval. Relevancy of information retrieved can also be improved by considering semantic relatedness between words which is a basic research area in fields like natural language processing, intelligent retrieval, document clustering and classification and word sense disambiguation. The web search engine based semantic relationship from huge web corpus can improve classification of documents. This paper proposes an approach for web document classification that exploits information, including both page count and snippets and also proposes the use of Artificial Bee Colony (ABC) algorithm as a new tool in the classification task. To identify the semantic relations between the query words, a lexical pattern extraction algorithm is applied on snippets. A sequential pattern clustering algorithm is used to form clusters of different documents. The page count based measures are combined with the clustered documents to define the features extracted from the documents. These features are used to train the ABC algorithm, in order to classify the web documents.

Keywords:- Artificial Bee Colony (ABC) algorithm, Document Classification, Term Document Frequency, Latent Semantic Indexing (LSI), Web Search Engine

1 Introduction

Classification is a form of data analysis that can be used to extract models describing important data classes. Such analysis can provide a better understanding of the data at large. Document classification can be applied as an information filtering tool and can be used to improve the retrieval results from a query process and to make good decisions. The documents to be classified may be texts, images, music etc. Each kind of document possesses its special classification problems. Documents may be classified according to their subjects or according to other attributes like document type, author and printing year. Mining useful information from a relatively unstructured source, such Hyper Text Markup Language (HTML), World Wide Web, news articles, digital libraries, online forums and other types of documents can be difficult. So extracting information from these resources and proper categorization and knowledge discovery is an important area for research.

Semantic similarity between terms changes over time and across domains. For example, *apple* is frequently associated with computers on the Web. This sense of apple is not listed in most general-purpose thesauri. A user, who searches for apple on the Web, may be interested in this sense of apple and not apple as a fruit. New words are constantly being created as well as new senses are assigned to existing words. Manually maintaining thesauri to capture these new words and senses is costly if not impossible. Each source of information provides a different viewpoint; a combination has the potential of having better knowledge than any single method.

Conventional document classification methods are directly performed in the entire document space. These conventional algorithms based on exhaustive searches of the document space become computationally infeasible. The self adaptability of population based evolutionary algorithms can be used to tackle the task of document classification. Artificial Bee Colony algorithm is considered new and widely used in searching for optimum solutions. This is due to its uniqueness in problem-solving method where the solution for a problem emerges

from intelligent behaviour of honeybee swarms. Thus in our approach, investigate the capability of ABC algorithm for web document classification using the features extracted from the selected dataset along with the features extracted from the web to improve the classification accuracy. The system forwards the user's query to a general-purpose internet search engine. Results are categorized based on the snippets for the user to choose from.

The reminder of the paper is organized as follows: section 2 discusses related work, section 3 explains the existing methodology, section 4 describes the proposed approach, section 5 presents the experimental results and section 6 concludes the paper with future work.

2 Related Work

Xiaogang Peng and Ben Choi [1], proposed to automatically classify documents based on the meanings of words and the relationships between groups of meanings or concepts. The bag-of-words document representation is simple, yet limited with two major problems. Word count cannot differentiate between related words in different documents or same words have different meanings under different context. Thus, rather than counting word occurrences, counting word senses might improve text classification by applying semantics to classification.

S. Doan and S. Horiguchi [2] proposed feature selection based on multi-criteria ranking of features. It performs better compare to conventional feature selection methods. But the limitation is how to fix the threshold for selecting the features.

Zakaria.E, Abdelattif.R and Mohd.Amine.B [3] have used WordNet concept to categorize text documents but the word sense disambiguation technique is not capable of determining the correct sense of words with multiple synonyms.

Gang Lu et al [4] discussed different Web search engines based word semantic similarity methods. Proposed a model called Revised CODC Model (RCODC) which uses snippets for improving accuracy of word similarity.

In [5], Jorge.G and Eduardo.M have explored the semantic relatedness measure between two words that use Web as knowledge source. Semantic relatedness measures quantify the degree in which words or concepts are related. Many semantic measures have been proposed in the past to compute degrees of relatedness among words, texts or concepts. In the Measures based on Thesauri and other lexical resources, Latent Semantic Analysis (LSA) is a statistical technique that leverages word co-occurrence from large unlabeled corpus of texts. But these methods result in a limited coverage. In the Measures based on Wikipedia, a method to represent the meaning of texts or words as weighted vectors of Wikipedia-based concepts using machine learning techniques is used. But wikipedia is still not comparable with the whole Web in the task of discovering and evaluation of implicit relationships. Measures based on the Web gives a guarantee of maximum coverage.

Arya.S and Lavanya.S [6] have proposed a similarity measure that combines various similarity scores based on page counts and lexico-syntactic patterns extracted from text snippets. The proposed work aims to classify the web documents which are most related to user's query into predefined classes or categories.

Aurangzeb et.al.[7] reviewed different machine learning algorithms for text-document classification like K-Nearest Neighbour, Decision Trees, Naive Bayes, Rocchio's Algorithm and Support Vector Machines. The authors concluded that the support vector machine classifier has been recognized as one of the most effective text classification methods and it also had highest classification precision.

The aim of feature-selection methods is the reduction of the dimensionality of the dataset by removing features that are considered irrelevant for the classification. Ikonomakis , Kotsiantis and Tampakas [8] discussed the different feature selection methods and text classification using machine learning.

Both machine learning and evolutionary computations have complementary strengths and limitations. All conventional evolutionary computations draw inspiration from the principles of Darwinian evolution and have been applied to a very wide range of optimization and search problems. An evolutionary algorithm deploys a randomized search. It is capable of searching through very complex problem spaces and get good results quickly for problems that change over time. It can also be used to reduce the processing time. The proposed methodology aims to include the classification of documents based on an evolutionary technique called the Artificial Bee Colony algorithm.

Yang [9] developed a virtual bee algorithm (VBA) to solve the numerical optimization problems. For optimizing multivariable numerical functions, Karaboga [10] has described a bee swarm algorithm called artificial bee colony algorithm which is different from virtual bee algorithm.

Mohd Afzi et.al [11] has applied the artificial bee colony algorithm for the first time as a new tool

for data mining particularly in classification tasks. The results obtained in the experiments indicate that ABC algorithm are competitive, not only with other evolutionary techniques, but also to industry standard algorithms such as PART, SOM, Naive Bayes, Classification Tree and K-Nearest Neighbor and can be considered as useful and accurate classifier. Basturk and Karaboga [12] compared the performance of ABC algorithm with the performance of genetic algorithm.

D.Karaboga and B.Akay [13] have made a comparative study of Artificial Bee Colony algorithm (ABC), Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Differential Evolutionary algorithm (DE) and Evolutionary Strategies (ES) for optimizing a large set of numerical test functions. While GA and DE employ crossover operators to produce new or candidate solutions, ABC algorithm does not. ABC algorithm produces the candidate solution from its parent by a simple operation based on taking the difference of randomly determined parts of the parent and a randomly chosen solution from the population. This process increases the convergence speed of search into a local minimum. In GA, DE and PSO the best solutions obtained is always kept in the population. However, in ABC, the best solutions discovered are not always held in the population since it might be replaced with a randomly produced solution by the scout bee. Apart from the maximum evaluation number and population size, a standard GA has three more control parameters like crossover rate, mutation rate and generation gap, a standard DE has at least two control parameters like crossover rate and scaling factor and a basic PSO has three control parameters like cognitive and social factors and inertia weight. The ABC algorithm has only one control parameter called limit. The performance of ABC algorithm is better than or similar to the other population-based algorithms with the advantage of employing fewer control parameters and it can be used efficiently for solving multimodel and multidimensional optimization problems.

3 Existing System

D.Bollegala et.al.[14], have proposed an automatic method to estimate the semantic similarity between words or entities in a query using web search engine for classifying them as synonymous or non-synonymous word pairs using support vector machine. Given two words P and Q, the problem of measuring the semantic similarity between P and Q is modelled as a function sim(P,Q) that returns a value in range of [0, 1]. If they are highly similar, sim(P,Q) will be close to 1. On the other hand, if they are not semantically similar, then sim(P,Q) will be close to 0. There are numerous features that express the similarity between P and Q using Page Counts and Snippets retrieved from a web search engine. Using this feature representation of words, the Support Vector Machine is trained to classify synonymous and non-synonymous word pairs.

Fig.1 illustrates an example of using the existing method [4] to compute the semantic similarity between two words. The steps are as follows:

1. Query a web search engine and retrieve page counts and snippets for input word-pairs from WordNet.
2. Calculate the word co-occurrences on web documents using either of the four measures namely WebJaccard, WebDice, WebOverlap or WebPMI.
3. The frequencies of the lexical patterns extracted from web snippets are calculated.
4. The lexical patterns that convey the same semantic relations are clustered together using a sequential pattern clustering algorithm.
5. Both page counts-based similarity scores and lexical pattern clusters are combined using support vector machine to find the semantic similarity measure.
6. The words are classified as synonymous or non-synonymous based on the similarity score.

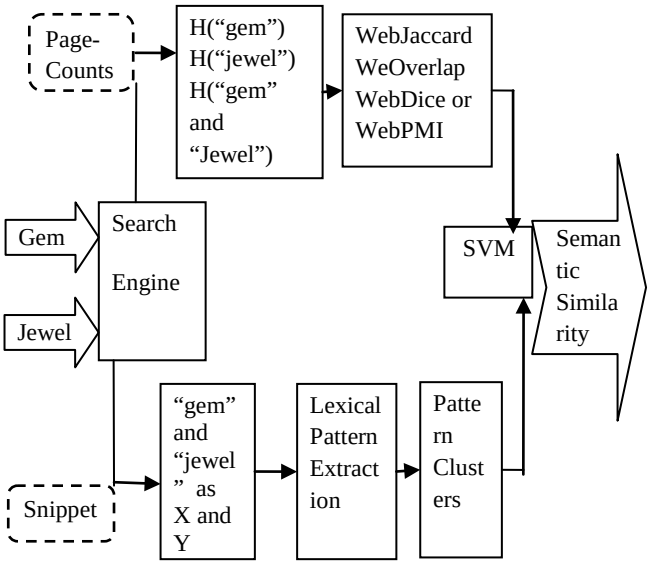

Fig.1 Outline of the Word-Pair Classification

D.Bollegala et.al.[14], does not involve in the classification of web documents into different categories based on the word pairs. In order to classify the web documents into different classes,

Kavitha.C, Sudha Sadasivam.G, and Kiruthika.S, [15] proposes an approach for web document classification that exploits information, including both page count and snippets. To identify the semantic relations between the query words, a lexical pattern extraction algorithm is applied on snippets. A sequential pattern clustering algorithm is used to form clusters of different patterns. The page count based measures are combined with the clustered patterns to define the features extracted from the word-pairs. These features along with the features extracted from the Reuters and 20newsgrroup data sets are used to train the Support Vector Machine, in order to classify the web document which is shown in Fig.2.

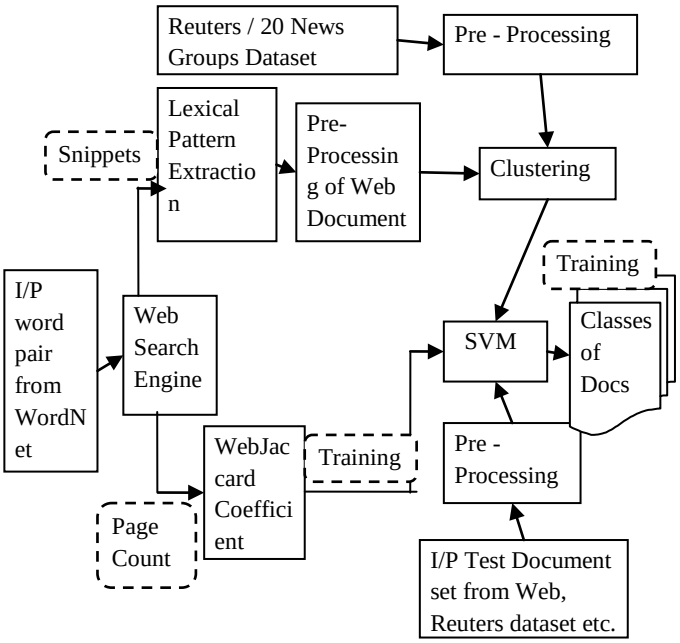

Fig. 2 Document Classification using SVM

Test documents from different corpus are tested to obtain the classification accuracy in terms of F1 measure. In order to improve the classification accuracy and processing speed, the proposed system classifies the documents based on an evolutionary technique called artificial bee colony algorithm and it is compared with the existing system.

4 Proposed System

The proposed methodology classifies the documents according to their content into categories. The proposed system architecture is shown in the Fig.3 WordNet [16], a manually created English dictionary, is used to generate the training data required for the proposed method. Around 2000 nouns are randomly selected from WordNet and a pair of synonymous words from a synset of each selected noun is extracted. These word pairs are given to the search engine [17] from which the page counts and the snippets are extracted.

The steps of the ABC based document classification approach are described as follows:

4.1 Similarity based on Page Count

The WebJaccard coefficient measure for page counts is defined as

$$WebJaccard\ (P,Q) = \frac{H(P \cap Q)}{H(P) + H(Q) - H(P \cap Q)} \quad (1)$$

where P and Q are two words in the query, H(P) and H(Q) denote the page counts for word P and Q respectively.

4.2 Extraction of Sub-sequences in a Snippet

The snippets are given to the lexical pattern extraction algorithm [14] to recognize the semantic relations that exist between two words. The sub sequences from the snippets are generated using the following conditions:

1. A subsequence must contain exactly one occurrence of each word P and Q.
2. The maximum length of a subsequence is *L* words.
3. A subsequence is allowed to skip one or more words. However, not more than *g* number of words consecutively.
4. All negation contractions must be expanded. For example, *didn't* is expanded to *did not*.

The frequency of occurrence of all sub sequences is counted and only those sub sequences that occur more than *T* times are used as lexical patterns. The web documents corresponding to the top ranked patterns are extracted. The parameters are set experimentally to $L = 7$, $g = 2$ and $T = 5$.

4.3 Document Pre-Processing and Feature Representation

The extracted web documents are pre-processed in order to transform the documents into a form suitable for automatic processing. The documents are represented using vector-space model. In this model, each document is represented as a vector $\vec{d}$. Each dimension in the vector $\vec{d}$ stands for a distinct term in the term space of the document collection. Each document is represented as a vector $\vec{d} = [w_1 w_2 \dots \dots w_n]$, where w_i is the term weight of the term t_i in one document.

The term weight value represents the significance of the term in a document. To calculate the term weight, the occurrence frequency of the term within a document and in the entire set of documents is considered. The weighting scheme combines the Term Frequency with Inverse Document Frequency (TF-IDF) [18, 19, 20].
The TF-IDF weighting scheme is used to ensure the effectiveness of document classification. The weight of term i in document j is given by

$$tf * idf_{i,j} = tf_{i,j} * idf_i \quad (2)$$

where term frequency is calculated as

$$tf_{i,j} = \frac{N_{i,j}}{NT_j} \quad (3)$$

$N_{i,j}$ is the number of times the term i appears in the document j and NT_j is the total number of terms in the document j. The inverse document frequency is calculated as:

$$idf_i = \log\left(\frac{|D|}{|d : t_i \in d|}\right) \quad (4)$$

where |D| is the total number of documents and $|d : t_i \in d|$ is the number of documents in which the term t_i appears. These TFIDF values and the list of documents are together represented as a vector space. The feature selection is employed to reduce the size of the feature space to an acceptable level in order to increase the overall performance [21].

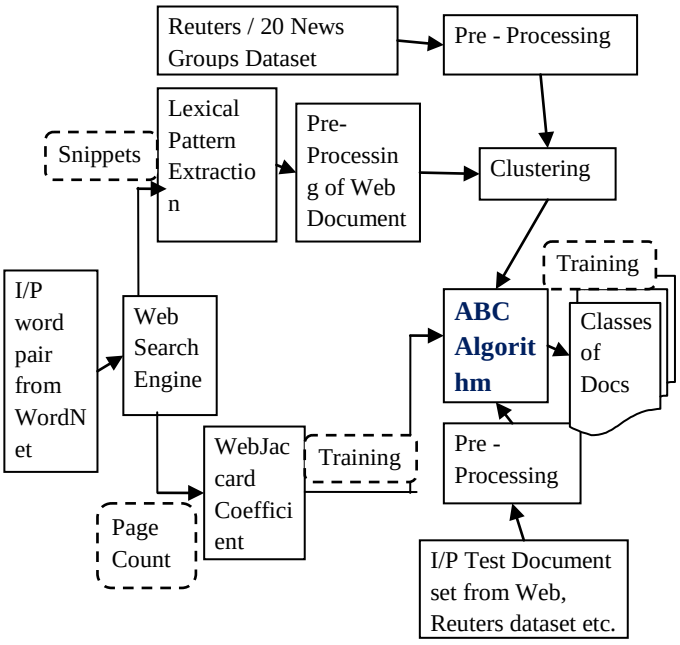

Fig.3 Outline of Proposed System

4.4 Clustering of Similar Documents

Cluster of similar documents are formed and labelled using latent semantic indexing (LSI). LSI analyzes the relationship between a set of documents and uses singular value decomposition (SVD) to find the semantic similarity between documents. LSI constructs a term-document matrix, *A*, to identify the *m* unique terms within a collection of *n* documents where each term is represented by a row and each document is represented by a column with each matrix cell initially representing the number of times the associated term appears in the indicated document. SVD is performed on the matrix [22, 23] to determine patterns in the relationships between the terms and concepts contained in the document. It computes the term and document vector spaces using the relation

$$A = TSD^T \quad (5)$$

where *T* = *m* by *r* term concept vector matrix; *S* = *r* by *r* singular value matrix;
D = *n* by *r* concept document vector matrix and *r* = rank of *A*.

LSI modifies the SVD to reduce the rank of *S* to size *k*, which effectively reduces the size of term and document vector matrix. This SVD reduction preserves the most important semantic information in the document and ignores the noise and other undesirable influences. This reduced set of matrices is denoted with a modified formula such as

$$A \approx A_k = T_k S_k D_k^T \quad (6)$$

The similarity of terms and documents within these vector spaces shows how close they are to each other. It is computed as a function of the angle between the corresponding vectors as

$$\cos\theta = \frac{A.B}{||A|| ||B||} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} * \sqrt{\sum_{i=1}^{n} (B_i^2)}} \quad (7)$$

4.5 Training the ABC algorithm

Training is the process of taking the content that belongs to specified classes and creating a classifier on the basis of that known content. The ABC algorithm is trained in two ways. First the page counts-based co-occurrence measures and the snippets-based lexical pattern clusters are combined into one feature vector and are used to train the ABC algorithm in combination with the Reuters [24] training dataset. Second the ABC algorithm is trained only with the Reuters dataset.

4.6 Classification using ABC algorithm

ABC algorithm is a new swarm intelligent algorithm [25] and consists of three essential components:

1. *Food Sources:* It represents a position of solution of the problem.
2. *Employed Foragers:* The number of employed bees is equal to the number of food sources. The employed bees store the food source information and share with others according to certain probability.
3. *Unemployed Foragers:* Their main task is exploring and exploiting food source. There are two choices for the unemployed foragers: (i) It becomes an *onlooker* and determines the nectar amount of food source after watching the waggle dances of employed bee and select food source according to profitability; (ii) It becomes a *scout* and randomly searches new food sources around the nest.

The preference of food source by an onlooker bee depends on the nectar amount $F(\theta)$ of that food source. In other words, the onlooker bee selects one of the food sources after making a comparison among the food sources. The probability with the food source located at θ_i that will be chosen by a bee can be expressed as

$$P_i = \frac{\sum F(\theta_i)}{\sum_{k=1}^{S} F(\theta_k)} \quad (8)$$

where $F(\theta_i)$ is the nectar amount present at the food source θ_i and S is the number of food sources around the bee hive. To evaluate the fitness value, the fitness function will be used for the classification instead of measuring the nectar amount. Its representation is defined as

$$F_1 = 2 * \frac{Precision \,.\, Recall}{Precision + Recall} \quad (9)$$

It considers both the precision and the recall to compute the score. Precision is the number of correct results divided by the number of all returned results and recall is the number of correct results divided by the number of results that should have been returned. The F1 score can be interpreted as a weighted average of the precision and recall.ABC algorithm is trained using the existing training set of Reuters dataset and web documents retrieved from top ranked snippets. The ABC model is typically developed through the training process. Finally, the results can be analysed by running the classifier on other contents and labelling them as belonging to one class.

5 Experimental Results

The major goal of document classification is to classify the documents relevant to user query. For the experiment, 1000 word pairs were taken from WordNet. Numerous patterns were extracted from the snippets. The web documents were retrieved for the patterns and clustered into many categories which were used for training the ABC algorithm.

Fig. 4 Comparison of No: of Categories formed in the Classification task

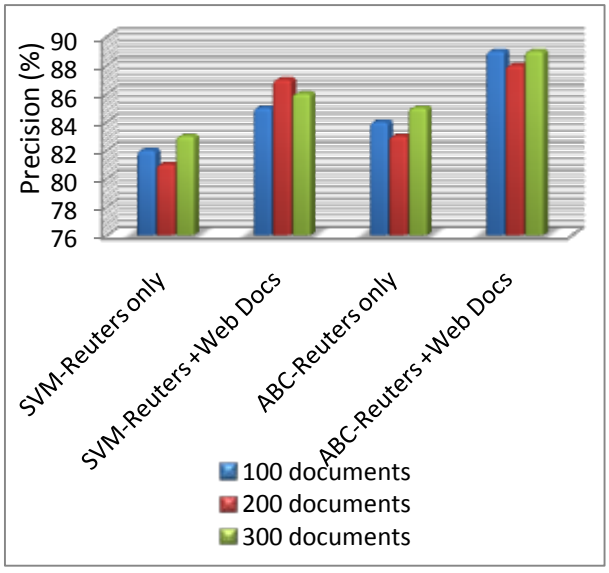

Fig.5 Comparison of Precision for Reuters dataset using Reuters dataset features alone and combination with web Documents features on SVM and ABC Classifiers

The ABC algorithm was also trained separately with the Reuters training dataset and also in

combination with the web documents and Reuters dataset. When the ABC algorithm was trained with only Reuters dataset, 10 large predefined classes were formed. When ABC algorithm was trained with a combination of both Reuters dataset and web documents, few other categories were formed like automobiles, business, organisation, electronics, places etc. The number of classes formed during the classification is shown in Fig. 4.The system was tested by giving different number of test documents from Reuters dataset and also web documents.

The experiments were conducted earlier with Support Vector Machine (SVM) algorithm and the performance of both ABC algorithm and SVM were compared. The standard performance measure for document classification is F1-Measure.

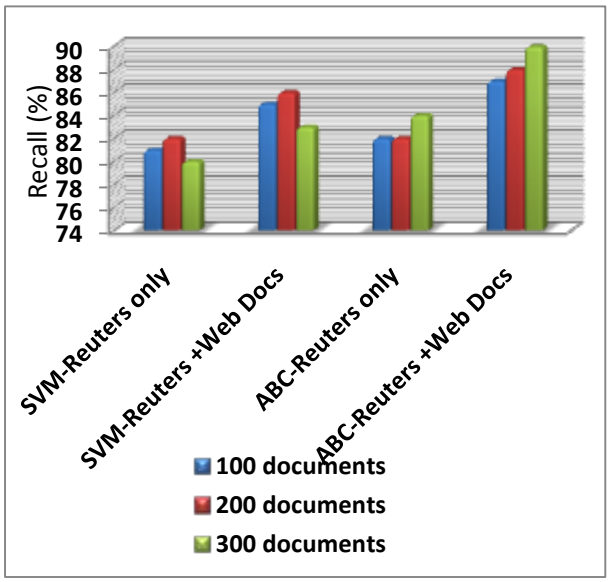

Fig.6 Comparison of Recall for Reuters dataset using Reuters dataset features alone and combination with web Documents features on SVM and ABC Classifiers

While comparing the classification of documents based on web documents and classification based on Reuters dataset, the results based on a combination of web documents and Reuters gave a performance increase as shown in the figures Fig.5 and Fig.6.

The F1 measure is calculated using equation 10

$$F_1 = 2 * \frac{Precision \,.\, Recall}{Precision + Recall} \qquad (10)$$

The F_1 score is a measure of a test's accuracy which is a balanced mean between precision and recall. Table 1 shows the classification accuracy for Reuters dataset in terms of F1 measure.

Table 1. F1 measure for Reuters Dataset

Features	Classifier	Average Precision	Average Recall	F1 Measure
From Reuter's dataset and Web documents	SVM	85	86	85%
	ABC	88	89	88%
From Reuters dataset alone	SVM	81	82	81%
	ABC	83	84	83%

The ABC algorithm was again trained with another dataset named 20News Groups. The experiment was repeated by giving 100, 200 and 300 input test documents for testing the performance of classification accuracy. Comparatively the classification based on ABC algorithm with a combination of web documents and 20News Groups gave a better performance than the SVM. The comparison of precision and recall of SVM and ABC algorithm for 20Newsgroup dataset is shown in the figures Fig.7 and Fig. 8.

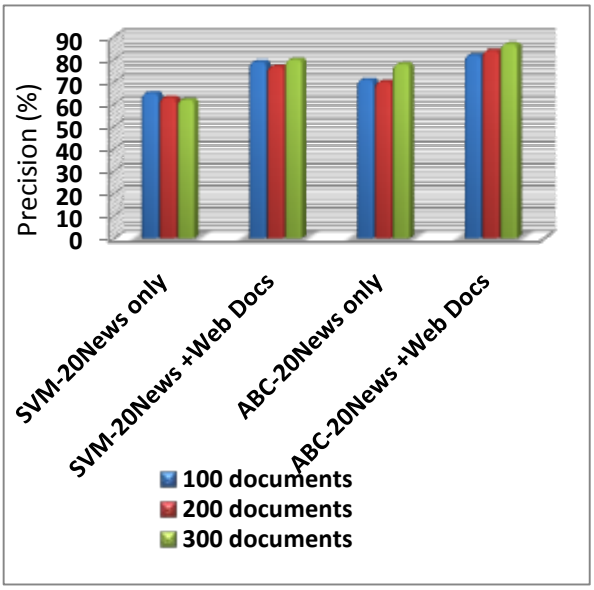

Fig.7 Comparison of Precision for 20Newsgroup dataset using 20Newsgrroup dataset features alone and combination with web Documents features on SVM and ABC Classifiers

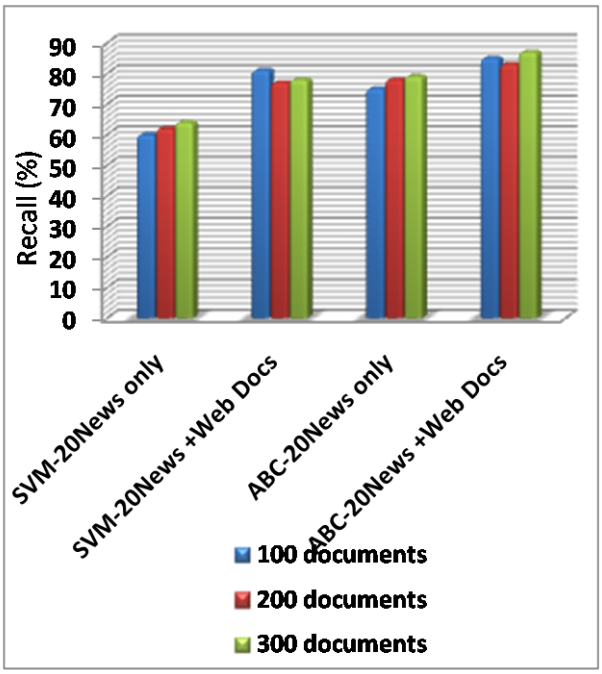

Fig.8 Comparison of Recall for 20Newsgroup dataset using 20Newsgroup dataset features alone and combination with web Documents features on SVM and ABC Classifiers

Table 2. F1 measure for 20NewsGroup Dataset

Features	Classifier	Average Precision	Average Recall	F1 Measure
From 20Newsgroup dataset and Web documents	SVM	78	79	78%
	ABC	84	85	84%
From 20Newsgroup dataset alone	SVM	63	62	62%
	ABC	73	77	75%

There was an increase in the percentage accuracy of classification when ABC algorithm was used, as the coverage of data was more in-depth. There was a decrease in the processing time when the documents were tested with ABC algorithm than with SVM.

6 CONCLUSIONS AND FUTURE WORK

Document classification is processed using artificial bee colony algorithm and support vector machine and the semantics is obtained by extracting the snippets and page counts from the web search engine for many pair of words. Training set is derived by using both the web search engine semantic and concept-based extraction using latent semantic indexing in order to retain the semantics among documents.

A comparison of training the ABC algorithm and SVM using Reuters dataset alone and with a combination of web documents and Reuters dataset has been carried out. The overall F1 measure for classification based on the proposed methodology using ABC algorithm is 88%. The F1 measure for classification based ABC algorithm for Reuters dataset using Reuters dataset feature alone is 83%. The experimental results indicate that the proposed method based on web documents yield better performance of precision and recall on unstructured documents due to the dynamic update of web contents and a thorough exploration of concepts. Moreover, the snippets provide the semantically related documents which are used to improve classification accuracy. The F1 for classification based on SVM for a combination of web documents and Reuters dataset is 85%. The F1measure for classification based on SVM for only features from Reuters dataset is 81%.

For 20Newsgroup dataset the F1 measure on SVM classifier with features from 20newsgroup and web documents is 78% and with ABC classifier is 84%. The results shows that the proposed methodology based on ABC algorithm gives improved performance compared to SVM as it involves fewer control parameters and it explores the concepts more thoroughly. The future work can include the parallelizing of clustering phase in order to reduce the processing time.

ACKNOWLEDGEMENT

Our thanks to Dr R Rudramoorthy, Principal, PSG College of Technology and Mr K Chidambaram Kollengode, Director, Cloud and Big Data Analysis, Nokia R&D, Bangalore, India, for their support. This project was carried out in Cloud and Big Data Analysis lab, PSG College of Technology, India.

REFERENCES

[1] Xiaogang .P, Ben .C, “Documents Classification Based on Word Semantic Hierachies”, *The IASTED International Conference on Artificial Intelligence and Applications*, 2005,pp.362-367.

[2] S. Doan and S. Horiguchi, "An Efficient Feature Selection using Multi-Criteria in Text Categorization for Naïve Bayes Classifier", *WSEAS Transactions on Information Science and Applications*, Vol.2, Issue 2, 2005, pp.98-103.

[3] Zakaria .E, Abdelattif .R, Mohd.Amine .B, "Using WordNet for Text Categorization", *The International Arab Journal of Information Technology*, vol.5, no.1, 2008, pp.16-24.
[4] Gang Lu, Peng Huang, Lijun He, Changyong Cu, Xiaobo Li, A new semantic similarity measuring method based on web search engines, WSEAS Transactions on Computers, v 9, n 1, 2010, pp 1-10.
[5] Jorge Gracia and Eduardo Mena, "Web-Based Measure of Semantic Relatedness", *Proc. of 9th International Conference on Web Information Systems Engineering, Springer*, 2008, pp.136-150.
[6] Arya .S.S, Lavanya .S, "An approach for measuring semantic similarity between words using SVM and LS-SVM", 2012, pp.1-4.
[7] Aurangzeb .K, Baharum .B, Lam Hong Lee, Khairullah .K, "A Review of Machine Learning Algorithms for Text-Documents Classification", *Journal of Advances in Information Technology*, vol.1, no.1, 2010, pp.4-20.
[8] M. Ikonomakis, S.Kotsiantis, V. Tampakas "Text Classification Using Machine Learning Techniques" *WSEAS TRANSACTIONS on COMPUTERS*, Issue 8, Volume 4, 2005, pp. 966-974.
[9] X.S. Yang, "Engineering Optimizations via Nature-Inspired Virtual Bee Algorithms", *Lecture Notes in Computer Science, 3562, Springer-Verlag*, 2005, pp. 317.
[10] D. Karaboga, "An Idea Based On Honey Bee Swarm for Numerical Optimization", *Technical Report-TR06*, Erciyes University, Engineering Faculty, Computer Engineering Department, 2005.
[11] M.A.Mohd Shukran, Yuk Ying Chung, Wei-Chang Yeh, A.M.Ahmad Zaidi and N.Wahid, "Artificial Bee Colony based Data Mining Algorithms for Classification Task", *Proc. Modern Applied Science*, 2011, vol. 5, no. 4, pp. 217-231.
[12] B. Basturk, D. Karaboga, "An Artificial Bee Colony (ABC) algorithm for numeric function Optimization", *IEEE Swarm Intelligence Symposium 2006*, May 12–14, 2006., Indianapolis, USA.
[13] D.Karaboga and B.Akay, "A Comparitive Study of Artificial Bee Colony Algorithm", *Elsevier Trans. Applied Mathematics and Computation*, 2009, pp.108-132.
[14] D.Bollegala, Y.Matsuo, and M.Ishizuka, "A Web Search Engine-Based Approach to Measure Semantic Similarity between Words", *IEEE Trans. Knowledge and Data Engineering*, 2011,vol. 23, no. 7, pp. 977-990,.
[15] Kavitha.C, Sudha Sadasivam, Kiruthika.S, "Semantic similarity based web document classification using support vector machine", *International Arab Journal of Information Technology* (IAJIT)(Accepted)
[16] WordNet-A Lexical Database for English, Princeton University. http://www.wordnet.princeton.edu/wordnet/
[17] Google Search, 2013. http://www.google.com/
[18] B. Everitt, "Cluster Analysis". 2nd Edition. Halsted Press, USA, 1980.
[19] G. Salton, "Automatic Text Processing", Addison-Wesley, 1989.
[20] Fouzi Harrag, Aboubekeur Hamdi-Cherif, Abdul Malik S Al-Salman and Eyas El-Qawasmeh," Evaluating the effectiveness of VSM model and topic segmentation in retrieving Arabic documents", *Journal of Computer Systems Science and Engineering*, 2011, vol.26, no.1.
[21] Kun Yue, Wei-Yi Liu, Li-Ping Zhou, "Automatic keyword extraction from documents based on multiple content-based measures", *Journal of Computer Systems Science and Engineering*, 2011, vol.26, no.2.
[22] Jianxiong .Y, Watada .J, "Decomposition of Term-Document Matrix Representation for Clustering Analysis", *IEEE Trans. International Conference of Fuzzy Systems (FUZZ*), 2012, pp.976-983.
[23] Muflikhah .L, Baharudin .B, "High Performance in Minimizing of Term-Document Matrix Representation for Document Clustering", *International Conference on Innovative Technologies in Intelligent systems and Industrial Applications*, 2009, pp.225-229.
[24] Reuters-21578 Text Categorization Collection, The UCI KDD Archive, Information and Computer Science, University of California, Irvine, http://www.kdd.ics.uci.edu/databases/Reuters21578
[25] D. Karaboga, B. Basturk, "On the performance of artificial bee colony (ABC) algorithm", *Elsevier Trans. Applied Soft Computing*, 2008, pp.687-697.