ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

Text Document Classification using Ant Colony Optimization and Genetic Algorithm

Monica Bali, Deipali Gore

Dept. of Computer Engineering, PES's MCOE, Savitribai Phule Pune University, Pune, Maharashtra, India

Assistant Professor, Dept. of Computer Engineering, PES's MCOE, Savitribai Phule Pune University, Pune, Maharashtra, India

ABSTRACT: The amount of information in digital form availablewith us is increasing rapidly day-to-day. The available informationwould be useful if we are able to access the relevantinformation efficiently. The main problem is to improve theefficiency and accuracy of text classification. To improve theefficiency of information we need to search, sort, index, store,and analyze the available information with the help of specifictools. In line with this one can read the texts and categorizethem manually when amount of information is less, but whatcan be done when information is in huge amount e.g. in termsof hundreds, and thousands of texts? To answer this we require a tool which uses a supervised learning task that assigns thepredefined category or class to new text documents. This initiatesthe need of some kind of automated application which works onthe text categorization. There are several algorithms used for textcategorization. In this paper we have proposed the new algorithmi.e. Ant Colony Algorithm for text document classification andcombination of ACO-GAfor feature selection. ACO provides theadvantages in providing the solution to discrete problems.

KEYWORDS: Text Categorization; Text Classification; Document Classification; Information Retrieval; Feature Extraction; Feature Selection; Ant Colony Optimization Algorithm; Genetic Algorithm.

I. INTRODUCTION

Automatic text categorization is an active research topicin the field of information retrieval and data mining, since the results are still subject to improvements. In general, textcategorization deals with a set of text documents and a set of categories. The aim is to prepare a computer applicationwhich can be able to assign a correct category to text documentdepending on its contents [1]. Text categorization is asupervised learning task that assigns the predefined categoryto new text document.

The text categorization classifies the text documents eitherinto only one category or into number of categories. In thispaper we are focusing on classification of text documentsinto only one category. The set of categories are predefined.The problem is to classify the texts based on their similarity.In text categorization documents are represented as featurevectors before the classification algorithm is applied on it.These features are mainly divided into two sets, as trainingset and test set. Features available in training set will be used in learning phase of algorithm to build the classifier. Then the classifier is used to classify the received text documents into predefined category. To do the estimation the classifier is applied on the test set and the result is estimated to see the performance of the classification [6].

Among too many methodologies, which are proposed for text categorization, Ant Colony Optimization (ACO) based method have involved a lot of attention. ACO has advantages in providingthe solution to solve complex optimization problems mainly in discrete optimization problem. ACO is encouraging algorithm in data classification and clustering [2]. Meta-heuristic optimization algorithm constructed on behavior of ants was introduced in the early 1990s by Dorigo and Caro(1999) [5]. ACO is newly developed branch of artificial intelligencecalled as swarm intelligence (SI). Swarm intelligenceincludes the learning of the emergent collective intelligenceof group of simple agents [4]. ACO is motivated by social behavior of ant colonies.

Originally ACO algorithm was developed for solving traveling salesman problem (TSP) and then has been successfully applied for graph coloring problem, routing in telecommunication, job-shop scheduling problem, etc. [2]. High dimensionality of feature space is another major problem in text categorization. Many more features of the original feature set are irrelevant to text categorization, which will increases the noise data and impact on the performance of classifier [5]. So we need to select the subset from the originalfeature set to reduce the dimensionality

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

of feature space which improves the efficiency and performance of classifier. There were several approaches applied to solve the problemof feature selection in text categorization. Yang and Pedersen(1997) have done a comparative study on five feature selectionmeasures for text categorization, which includes documentfrequency (DF), Information gain (IG), mutual information(MI), aχ2 test (CHI) and term strength (TS) and foundχ2 test (CHI) and information gain provides more effectivefor optimizing classification results, and document frequencyis a better option for efficiency and scalability if a small degradation in effectiveness is affordable [5]. Among all the approaches which are proposed for feature selection, genetic algorithm (GA) which is population-based method and ant colony optimization (ACO) - based method have more attention.In our proposed approach we are considering combined ACO-GA method for feature selection which selects the most representative features for text categorization.

II. RELATED WORK

A. *Overview of Text Categorization:*

Thegoal of automatic text categorization is to classify atext document into the correct category depending upon itscontents; the category states to the subject or class [1]. The setof categories are predetermined. The main problem is to groupthe text documents by their similarity. In text categorization theclassification is nothing but assigning the predefined categoryto the new text documents based on the likeness of the textto a category, since the category is thoroughly related to themeaning of the text.

To recognize the correct category associated to a text, thefollowing steps are essential [1]:

1) Learning steps includes

a. The class for every text.

b. From the corpus available with us we select the k descriptors (*t*1 … *tk*) which are most relevant in problem solving.

c. A table of descriptors X individuals and its values for every word of text.

2) The classification of new word for new text d_x includes two stages

a. Weighting the occurrences *t*1... *tk* of terms in text to classify d_x.

b. Implementation of a learning algorithm on these occurrences and the previous table to predict the labels of the text d_x.

B. *Ant Colony Optimization:*

Ant Colony Optimization is a meta-heuristic for the solutionof hard combinational optimization (CO) problems inspiredby nature-ants, which was first introduced by M. Dorigo andhis colleges in early 1990s [6]. The idea was motivated bythe foraging behaviour or real ant colonies. First algorithmintroduced and applied for solving the travelling salesmanproblem and then applied on many other problems such asquadratic assignment problem, network routing, scheduling, etc. [6].

The first ACO algorithm developed was the Ant System(AS) (Dorigo, 1992) [6] and then numerous improvement ofthe AS have been developed. ACO algorithm is based on acomputational paradigm motivated by real ant colonies.

ACO is constructed after the collective searching behaviour of nature ant colony. While moving, the ants will leavesome secretion (pheromone) produced by the smell to transferinformation and complete the tasks by cooperating with eachother. The ants in colonies will follow the odoriferous roadand leave the pheromone while moving. Until all ants willselect the shortest path the ants in this path will leave morepheromone on the road. Relation between TSP problem andtext categorization using ACO is as follows.

1) For construction of graph, the node represents documents.

2) The pheromone i.e. distance between two cities(nodes) may be considered as similarity between documents. Thesimilarity (cosine similarity) is calculated using following formula

$$S(i,j) = \cos(i,j) = \frac{\sum w_i \sum w_j}{\sum w_i^2 \sum w_j^2}$$

The optimal path of every category will be found after thepredefined no of ants crawl in this manner. The path withhighest pheromone value is selected by comparing. The categoryassigned to the path is assigned to the text.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

C. *Genetic Algorithm*

Genetic algorithm is population-based search algorithm inspiredfrom principles of natural evolution called as evolutionaryalgorithm (EA) [4]. These algorithms are general-purposeoptimization algorithms with probabilistic components [5].

Originally, GA was developed to optimally solve sequentialdecision processes but over the years, it has been used inlearning as well as optimization problem. GA works with apopulation of points instead of working with a single point.Each point is considered as vector in hyperspace. Each vectoris called a chromosome which is represented as a binarystring. Number of elements in each chromosome is same asthe number of parameters in optimization problem. A generalseries of operations carried out while applying a GA are asfollows,

- Initialize the population.
- Calculate fitness for each chromosome in population.
- Reproduce the selected chromosome to form a newpopulation.
- Apply crossover and mutation on the population.
- Repeat from second condition until some condition is met.

III. PROPOPSED ALGORITHM

In the literature we find that the commonly used classificationalgorithm is quite slow in terms of processing time as well as the classification rate. In our proposed systemwe are applying combination of ACO and GA for featureselection which will give more precise subset of feature fortext classification.To determine the text documentcategory, we adopt thealgorithm of ant colony optimization (ACO) suggested in [2].Our choice is inspired by the flexibility of the meta-heuristics which makes possible its application that are common to NPhard. In general, a text categorization system includes numerousessential like feature extraction and feature selection. Once the text documents are pre-processed, feature extraction isapplied to convert the input text document into a feature vector. For dimensionality reduction feature selection is applied to thefeature vector.The overall process of proposed system is depicted in Fig.1 and explained in the following steps.

A. *Document preprocessing*

The document set provided as an input must be pre-processedas it contains texts that are irrelevant for classification. The pre-processing includes tokenization, stop word removal,stemming and term weighting.

1) Tokenization: It is the process of splitting stream oftext into meaningful words or phrases. The words are splitbased on the special delaminating characters such as spaces,punctuation, and symbols etc.

2) Stop Word Removal: Frequently occurred words, likepronouns, prepositions and conjunctions in English e.g. ‘it’,‘in’, ‘and’, etc. are known as stop words. These words fromthe text documents are having a very low discriminative value.It includes creating a list of stop words and then scanning thetokens to remove the stop words occurred. These common stopwords are collected from [9]. The final list of distinct wordsincludes 6508 words.

3) Stemming: It is the process of finding the root wordof the token. For example, the words “purification”, “purity”,“purify”and “purifying”having stemmed root as “pure”. Stemmingwords helps to reduce the dimensionality of the featurespace. The Porter stemming algorithm is used, which is anatural language processing (NLP) by removing the suffix[10], to narrow down the size of the feature space.

4) Term Weighting: TF-IDF is a term weighting approachwhich is one of the widely used methods to evaluate theimportance of a term in the corpus or identifies how relevanta term is to the classification. It can be calculated as formula(2) as follows.

$$W(i,j) = tf_{ij} \cdot \frac{N}{df_{ij}}$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

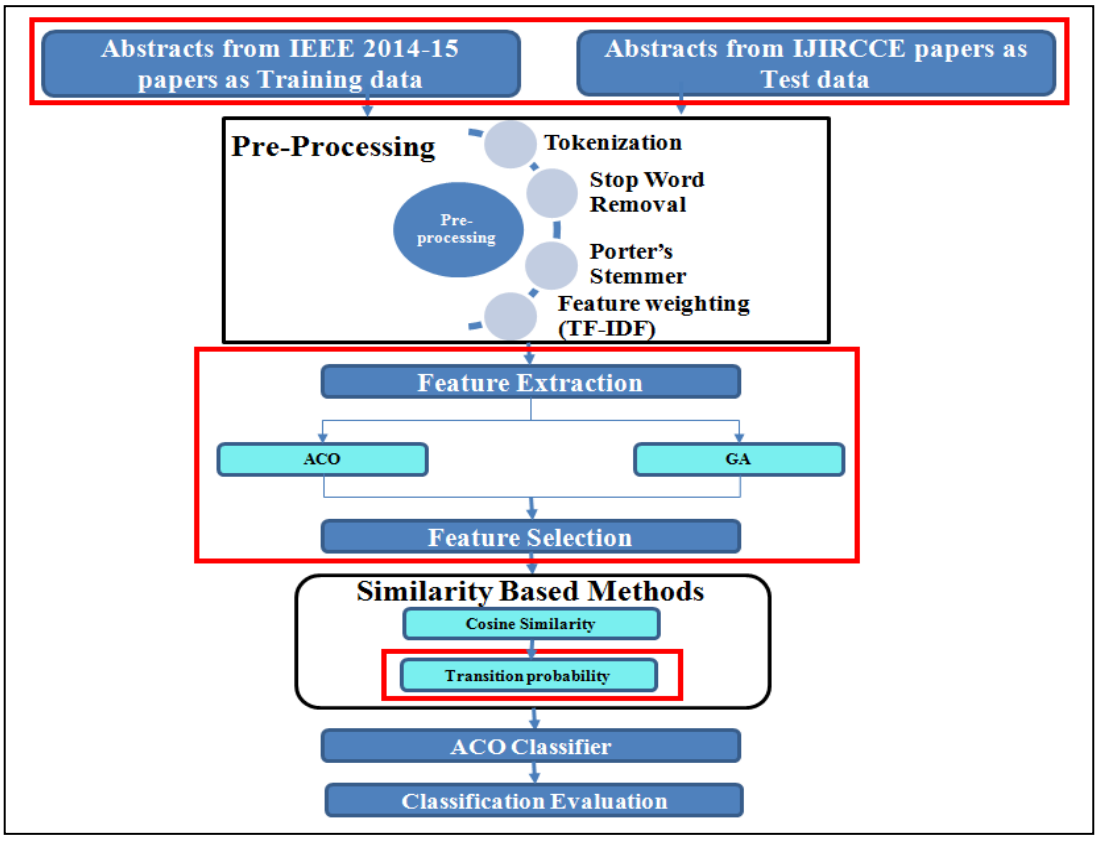

Fig.1.Proposed System Architecture

B. *Feature Extraction*

It is the process of converting the text feature into featurevector. For the representation of text we are going to use thevector space model in our proposed system.

C. *Feature Selection*

Feature selection is used for dimensionality reduction oforiginal feature set to get the more relevant feature space forclassification. In our proposed system we used thecombination of ACO and GA proposed in [4].

D. *Similarity Based techniques*

1) Cosine Similarity: The similarity between two documentswhich are considered as nodes can be calculated usingcosine similarity. The cosine similarity is calculated usingformula (1).

2) Transition probability: The transition probability can becalculated using formula (3) as follows.

$$P_{ij} = \frac{\tau_j}{\sum_{\tau} \tau_i}$$

The next node for the classification will be selected by takingproduct of formula (1) and formula (3).

E. *ACO classification*

In the proposed system we adopt the algorithm of Ant Colony Optimization (ACO) suggested in [2]. Once we get theprécised set of features, which is outcome of proposed ACO-GAapproach for feature selection. On this feature set we will apply the ACO based approach for text categorization.There were several techniques available for text classification.The proposed approach will give the better results forclassification of text documents into its correct category thanother approaches. The proposed approach can improve theefficiency and performance of the classification.

F. *Evaluation Of Classification*

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

To evaluate the performance of the classifier is evaluated according to the accuracy results. In order to compare the predicted categories assigned by classifier with the actual categories of the test documents, first of all the number of True Positives, False Negatives and False Positives are determined, then precision, recall and accuracy is computed using these values.

IV. MATHEMATICAL MODEL

Mathematical Model for proposed approach is as follows:

Let S, be the proposed system which can be represented as

$$S = \{\{I\}, \{P\}, \{O\}\}$$

Where,
I ->Input document collection (for Training and Testing)
P -> Functions used
O -> Test document labeled with the appropriate Domain
Where,

$$P = \{ f1, f2, f3, f4 \}$$

f1 ->Term Weighting (TF-IDF)
f2 -> Feature Selection Method (ACO-GA)
f3 -> Similarity Based Methods (Cosine Similarity and Transition Probability)
f4 -> Evaluation Parameters (Precision, Recall and Accuracy)

V. EXPERIMENTAL SETUP AND RESULTS

A. *Dataset Used*

To evaluate the effectiveness of the text document classificationalgorithms, several text collection available. These collectionsare useful for research in Information Retrieval, Natural Languageprocessing and other corpus-based research.For evaluating the algorithmic approach that we introducein this paper, we have selected the training dataset as:IEEE Online Papers[11]: The IEEE papers published in year2014-2015 and test dataset as: IJIRCCE online papers [12]. From the collection of input papers we are extracting the 'abstract'and providing as aninput to the system. Pre-processing steps applied on the abstractextracted from the papers and store in the database. And then further steps from proposed system applied on this collectionfor document classification to classify the document depending upon the content of theabstract extracted. For classification of text documents, we considered the 10 domains as categories.Table I shows the ten categories along with the number of training and test examples in each.

Domain Name	Total No. of Training Documents	Total No. of Test Documents
Cloud Computing	19	30
Data Mining	11	27
Database Security	12	6
Distributed System	14	10
Genetic Algorithm	8	5
Image Processing	7	3
Mobile Computing	7	6
Software Application	5	6
VANET	7	2
Wireless Sensor Network	7	2

Table I: No of Training and Test Documents

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

B. *Performance Measure*

To evaluate the performance of the classifier theimportant tool used is the confusion matrix. Confusion matrixis helpful in pin pointing the opportunities to improve theaccuracy of the system.

	Document belonging to the Category	Document not belonging to the Category
Category assigned to the document by the classifier	TP_i	FP_i
Document category rejected by the classifier	FN_i	TN_i

Table II: Confusion Matrix

In information retrieval systems precision, recall and accuracy arethe most used measurements to evaluate the performance.According to the Table II, precision and recall are defined asfollows,

$$Precision = \frac{TP_i}{TP_i + FP_i}$$

$$Recall = \frac{TP_i}{TP_i + FN_i}$$

$$Accuracy = \frac{TP_i + FN_i}{N}$$

C. *Results*

To show the utility of proposed system we compare proposed algorithm with ACO-based algorithm proposed in [1]. Various values were tested for the parameters of proposed algorithm. The results show that the highest performance is achieved by setting the parameters to values shown in Table III.

Method Used	Iterations/ Generations	Initial Pheromone	Crossover	Mutation
ACO	User Defined	2	-	-
GA	User Defined	-	0.84	0.72

Table III: ACO and GA Parameter Setting

For analysing the performance of ACO and ACO-GA algorithms we have considered following confusion matrix for test documents.

	No. of Documents Belonging to the Category		No. of Document not belonging to the Category	
	ACO	ACO-GA	ACO	ACO-GA
Category assignedd to the document by the Classifier	63	72	5	3
Document Category rejected by the Clasiifier	3	2	18	20

Table IV. Confusion Matrix for Document Classification using ACO and ACO-GA

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

Analysing the precision, recall and accuracy shown in Table V, we see that on average, ACO-GA algorithms obtained a higher accuracy value than ACO.

Method Used	Precision (%)	Recall (%)	Accuracy (%)
ACO with Cosine Similarity	94.02	96.92	67
ACO-GA with Cosine Similarity and Transition Probability	96	97.3	76.29

Table V. Comparative Results of Document Classification using ACO and ACO-GA

The results show that as the percentage of selected features exceeds 9% in accuracymeasures, the ACO-GA algorithm outperforms ACO algorithm.

To graphically illustrate the progress of the ant colony as it Searches for optimal solutions,we take category as the horizontal coordinate and the no of documents classified into thecategory as the vertical coordinate. This should illustrate the process of improvement ofthe best ant as the number of features increase.

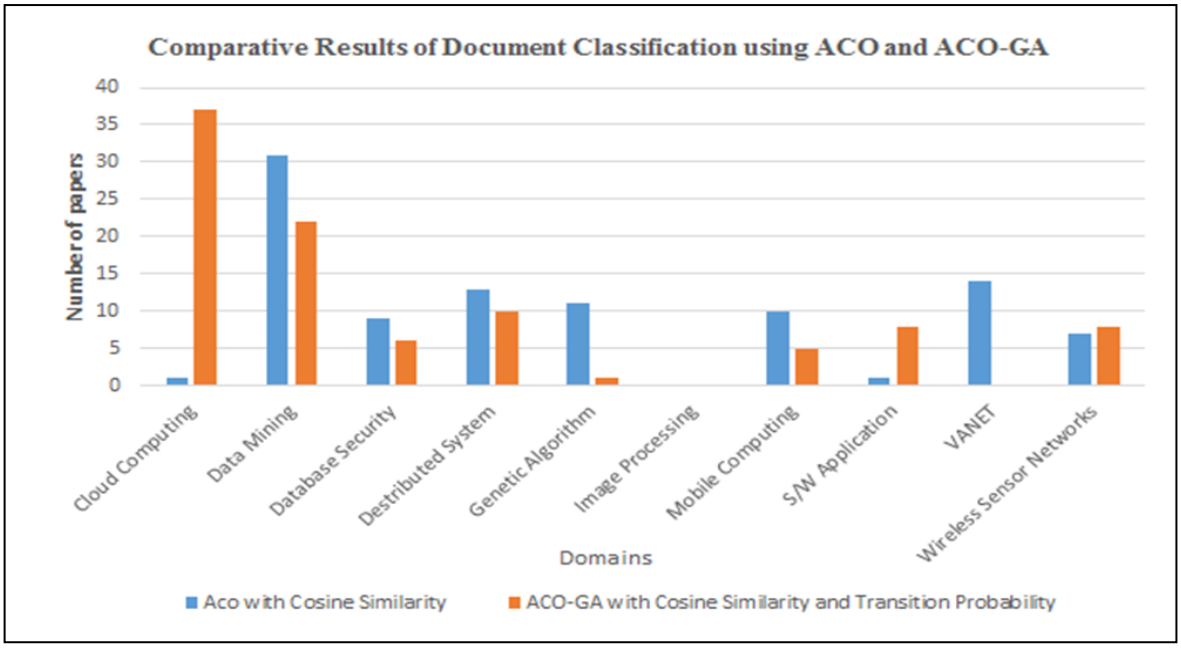

Fig.2. Comparative Results of Document Classification using ACO and ACO-GA

VI. CONCLUSION

In this paper an ant colony optimization / geneticalgorithm hybrid feature selection algorithm for text classification is presented. Inthe proposed algorithm, the classifier performance and thelength of selected feature subset are adopted as heuristicinformation. Therefore, it can select the optimal featuresubset without the prior knowledge of features.

Proposed algorithm has the ability to converge quickly; it has a strong search capability intheproblemspace and can efficiently find minimal feature subset.Experimental results demonstrate competitive performance.Proposed algorithm, ACO-GA, was compared with anexisting ACO-based method in text categorization. In order toevaluate the performance of proposed algorithm,experiments were carried out on the dataset in the literature, i.e. IJIRCCE papers.The computational results indicate that proposedalgorithm outperforms ACO, since it achieved better performance with the lowernumber of features.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

The combination of ACO-GA increases the Precision and Recall which results in moreaccurate subset selection for categorization and improves accuracy and computing timeof text document categorization.The new document is easily classified into its belonging category within less amount oftime.

REFERENCES

1. Nadia Lachetar and Halima Bahi,'Application of an Ant Colony Algorithm for Text Indexing', Computer Science department,IEEE, 2010.
2. Lijiuan JIAO and Liping FENG, 'Text classification based on Ant ColonyOptimization', Department of Computer Science, Xinzhou Teachers University,Xinzhou, China, IEEE pp. 229-232, 2010.
3. M. H. Aghadam, N. G. Aghaee and M. E. Basiri, 'Application of Ant Colony Optimization for Feature Selection in Text Categorization',Congress on Evolutionary Computation, IEEE, pp. 2872-2878, 2008.
4. M. E. Basiri and S. Nemati, 'A Novel Hybrid ACO-GA Algorithm for TextFeature Selection', Congress on Evolutionary Computation, IEEE, pp. 2561-2568, 2009.
5. M. H. Aghadam, N. G. Aghaee and M. E. Basiri, 'Text feature selection using ant colony optimization', Elsevier, 2008.
6. M. F. Zaiyadi and B. Baharudin, 'A Proposed Hybrid Approach forFeature Selection in Text Document Categorization', World Academy ofScience, Engineering and Technology, Vol. 4, 2010.
7. Shruti Tiwari and Prof. Anurag Jain, 'Selecting Feature Using Ant ColonyOptimization Algorithm, International Journal of Electrical', Electronics and Computer Engineering, pp. 206-211, 2014.
8. R. Jensi1 and Dr.G.Wiselin Jiji2, 'Survey on Optimization Approaches to Text Document Clustering', International Journal on Computational Science & Applications, Vol.3, No.6, 2013.
9. Webconfs Website (2006) Stop words list. Retrieved February 26, 2006,from http://www.webconfs.com/stop-words.php.
10. Rijsbergen, C.J.V., Robertson, S. E. and Porter, M. F. (1980) New modelsin probabilistic informationretrieval, London: British Library. (BritishLibrary Research and Development Report, no. 5587).
11. IEEE IEL online, (2014-2015) http/ieeexplore.ieee.org/.
12. IJIRCCE online http://www.ijircce.com/.

BIOGRAPHY

Monica Ramling Bali received the B.E degree in Information Technology Engineering from T. P. C. T's College of Engineering in 2011 from BAMU, Aurangabad. She is now pursuing M.E. degree in Computer Engineering from P.E.S.'s Modern College of Engineering, Savitribai Phule Pune University.