iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 Issue: IX Month of publication: September 2016

DOI:

Volume 4 Issue IX, September 2016
ISSN: 2321-9653

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Comparison of Performance in Text Mining Using Text Categorization of Semi Structured Data

M. Nandhiya[1], Ms. M. Sakthi[2] M. Sc, M. Phil, Ph. D
[1]*M. Phil Scholar, Department of Computer Applications, N.G.M., College, Pollachi, India.*
[2]*Assistant Professor, Department of Computer Science, N.G. Mcollege, Pollachi, India.*

Abstract -Text mining or knowledge discovery is that sub process of data mining, which is widely being used to discover hidden patterns and significant information from the huge amount of unstructured data. The enormous amount of information stored in unstructured / semi structured data cannot simply be used for further processing by computers, which typically handle text as simple sequences of character strings. Therefore, specific pre-processing methods and algorithms are required in order to extract useful patterns. In this study, we compared the performance of these classifications by applying the method of Bayesian methods, k-NN, decision trees, SVM, and as a neural network in classification on famous 20_newsgroup dataset from CMU Text Learning Group Data Archives, which has a collection of 20,000 messages, collected from 20 different net news newsgroups. The news will be classified according to their contents.

Keywords - Text Mining, Categorization, Natural Language Processing, Machine Learning, Bayesian, k-NN, SVM, Decision tree.

I. INTRODUCTION

Text mining is one type of data mining technique. Text mining discovers the previously unknown information extracting it automatically from different source. Text mining is similar to data mining. But the data mining dealing with structure data and text mining dealing with unstructured or semi structure data. Like email, text document and etc., in a text mining main goal is to discover the previously unknown information. The problem is that the result is not relevant to users need.

Discovering patterns is a great challenge due to huge amount information increases day by day to find accurate knowledge [1]. Text mining is used to perform this task which gives better performance for finding the relevant information. To resolve this problem, various techniques of text mining are discussed in this work. Text mining is nothing but extracting patterns from number of unstructured text documents. This technique is also called as Knowledge Discovery from Text (KDT).Text documents are considered as semi-structured or unstructured format. Computer has not that much capability to easily differentiate linguistic patterns as compare to human. But the computer can process text at high speed and in large volume. So, text mining becomes useful for computer to examine unstructured data. This technique employs numbers of algorithms to for converting unstructured text into useful patterns. Text Summarization, text categorization and text clustering these are the functions of text mining [2].

Text mining process starts with the collection of document from different source. Text mining tools help to retrieve a document and perform preprocessing on it. Then document go to next stage it apply text mining techniques like classification, clustering, visualization, summarization, and information extraction. And the last step analyzes the output data. For analyzing the output of text the users could navigate through in order to achieve the perspective [3] based on following figure 1.

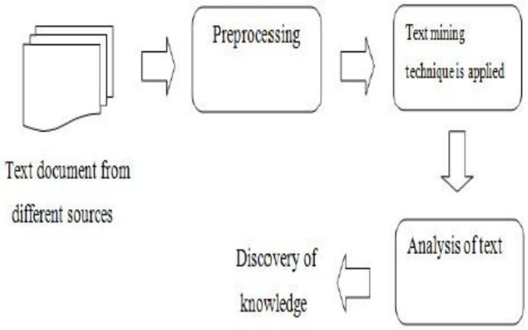

Text document from different sources
Preprocessing
Text mining technique is applied
Analysis of text
Discovery of knowledge

Fig 1.Text mining frame work

This paper provides the general overview of text mining, text classification and its techniques and classification on dataset.

II. LITERATURE SURVEY

In [5] S. Subbaiah illustrated how to extract knowledge from large text document. My initial study shows that they proposed system which uses ODP taxonomy and domain ontology and dataset to cluster and identify the category of text document. Here they use probabilistic classifier (Naive Bayes classification) for text mining from text document. Proposed work is based on three step 1) Pre processing which pre process on input text document and removed stemmed, stop words, and split into paragraph and statement. 2) Rule generation here it generate positive and negative rule. 3) Probability calculation and generated positive and negative rule is used to calculate the probability value. According to probability value each term set or pattern are identified from text document. Based on probability value sort the positive and negative probability value and select the category from most top probability value. In this paper with the help of probabilistic classifier its generate good result but its have little false indexing. Here they used Reuter's data set and each corpus data split in to ten categories. They use 70% training data set and 30 % testing set. In future we generate a effective rule and change in probability calculation to improve the overall result of text mining.

III. PREVIOUS WORKS

Fabric Colas and PavelBrazdil [4] compared to SVM, both K-NN and naive Bayes are very simple and well understood. SVM is however; more appealing theoretically and in practice, its strength is its power to address non-linear classification tasks. Unfortunately, most of the tasks examined here were not like that. The simplest SVM based on a linear kernel and a large error was found to be sufficient. Regards k-NN, the optimal number k of nearest neighbours is interestingly close to the ones used in other comparative studies carried out on different problems.

Lee Junyeon, Shin Seungsoo and Kim Jungju [5], compared the performance of these classifications by applying the method of Bayesian methods, k-NN, decision trees, SVM, and as a neural network in classification of unstructured newspaper article into given categories. In the experiment result, the SVM model has a high F-measure value relative to other models, and has shown stable results in the classification information and recall rate. Also, this model showed a high F-measure value in the classification of a more granular list. The methods of k- nn and decision tree show slightly lower performance than SVM, they are turned out to be appropriate models using classification problem cause of having advantages to easy interpretation and short learning time.

Shuzlina Abdul-Rahman, SofianitaMutalib, NurAmiraKhanafi [6] Describe that finding the content from large text document is time consuming. Text categorization is process to assign a text in to pre define categories. The paper explores several feature selection that use to reduce dimension and feature space. The support vector machine adapt here and its fast and perform well. The accuracy is higher in feature selection, and ability to handle categorization problem for large data set

IV. TEXT MINING

The purpose of text mining is to find useful patterns from the large documents and it means to apply an algorithm to the method of the text data and the statistical machine learning. In the aspect of finding a pattern and extracting the information from a big data, text mining is similar to data mining, except using unformulated data.

A. Text Mining Process

Overall process of text mining is depicted in the figure 1. Text mining process comprises of

1) Text pre-processing.
2) Text Transformation.
3) Feature selection.
4) Pattern discovery and
5) Evaluation.

B. Text Pre-processing

Text pre-processing is the initial step of text mining which reads one text document at time and processes it. This step divides into following main three subtasks-

1) Tokenization.
2) Stop Word Removing.
3) Stemming.

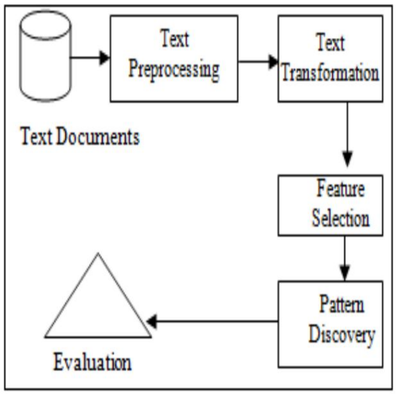


Fig 2: Text Mining Process

a) *Tokenization:* Generally text document contains multiple sentences. So this process divides whole sentence into words by removing comma, spaces, punctuations etc.

b) *Stop Word Removing:* This process removes stop words such as “the”, “are”, “a” or any tags like HTML tag etc.

c) *Stemming:* Stemming is applied after stop word removal by reducing the word to its root word. E.g. “playing”, “played” are stemmed to “play”.

C. Text Transformation

Text transformation has the role of conversion of text document into words so that it will useful for further processing.

D. Feature Selection

It performs removing features that are considered unrelated for mining purpose.

E. Pattern Discovery

Pattern discovery is one of the important processes that use methods for discovering patterns. Methods include clustering, classification, summarization, information retrieval, topic extraction etc.

V. CLASSIFICATION TECHNIQUES

Text classification (categorization) is the procedure of assigning a category label to documents. In tradition, decision about the label assignment is based on information gained by using a set of pre-classified text documents in order to build the classification function. So far, many different classification techniques have been proposed byresearchers, e.g. naïve Bayesian method, support vector machines (SVM), Rocchio classifier (Rocchio, 1971) (vector space representation),k-NN, decision trees, neural networks and many others.

A. Bayesian Method

The Bayesian method is one of the most widely used algorithms in the field of document classification, calculate the posterior probability that it will be assigned to each category receives a single document using the Bayesian theory, such as the equation (4.1).

$$P(C_j \mid d) = \frac{P(d \mid C_j)P(C_j)}{P(d)} \qquad (4.1)$$

In this Formula, d means random document, C_j means j-th category. P(d) has the same value to all categories, doesn’t need to calculate probability. Bayesian method assumes all the words are independent from each other, assignment of the category that occurs in the document is that mutually exclusive.

So, the *P(Cj|d)* can be calculated as follow equation (4.2)

$$P(C_j \mid d) = P(C_j)\prod_{i=1}^{n} P(w_i \mid C_j) \qquad (4.2)$$

In (4.2), *P(wi|Cj)* means (the number of *wi*'s occurrence in *Cj*)/(the number of all words' occurrence in *Cj*), and *P(Cj)* means (the number of documents allocated in *Cj*)/(the number of all documents).

The Baysian method calculates the probability of being classified in each category by assigning a document to a category having the maximum value, because a number of categories are categories to calculate the conditional probability for each of the categories having the highest probability belong to the document.

B. K-Nearest Neighbor

The k-NN classification method is to select one pattern from among the stored pattern with a distance of at least the learning data with respect to any particular pattern and classifies the category number which belongs to the category of the given pattern.

The k-NN algorithm is primarily used to the degree of similarity between the new document (*dx*) and the learning document (*dj*) to find the degree of similarity is high top k neighbouring documents in document study groups. Representative degree of similarity is the cosine similarity degree is calculated as in equation (4.3).

$$sim(d_x,d_j)=\frac{\sum_k t_{xk}\times t_{jk}}{\sqrt{\sum_k (\mathrm{t}_{xk})^2}\times\sqrt{\sum_k (t_{jk})^2}} \qquad (4.3.)$$

In the equation (4.3), t_{xk}, t_{jk} means the weight value of word *k* appeared in d_x, d_j.

The degree of similarity between the new document and document neighbours is used as the document category, the weight of the neighbourhood, if neighbouring pages when the category weights, share category is high.

When extracting k learning documents among the new document and the order of similarity using the cosine similarity, the category assigned to each k learning documents is a candidate list of categories to be assigned to the new document. The order to find the best category to be assigned to the new document category from the candidate list, calculate the compliance with total frequency of each category or the similarity score per category such as expression (4.4) in the learning document classification.

$$\mathrm{rel}(C_k\,|\,d_x)\approx\sum_j sim(d_x,d_j)\times\{(C_k\,|\,d_j) \qquad (4.4.)$$

In the k-NN algorithm selection of k it will have a large impact on the classification result. Likely k is not too small, there is a possibility of overcharging as the sum of the noise data for training, as opposed to the data groups classified as near to that k is too large classification exist.

C. Decision Tree

Decision tree is made a decision rule to classify the function. In supervised learning problem it is sometimes important to the prediction of the final model and analysis even more emphasis on predictive analysis or according to circumstances. By dividing the regions of each variable repeatedly with supervised learning techniques, decision tree creates a rule with incorrect prediction and correct analysis relatively to the other supervised learning techniques for the whole area.Rules created by the decision tree has the advantage of being easy to understand and easy to implement, because they have if-then structure.The analysis of the decision tree is composed of growth of the decision tree, pruning, feasibility study, and analysis and forecasting. The most widely used in decision tree algorithm is CHAID. This can be applied to all types of the target and classification variables, such as nominal, the order-type, continuous type.

Advantages

1) Decision trees are simple to understand and interpret.
2) It requires little data and are able to handle both numerical and categorical data.
3) We validate this model using statistical tests.

D. SVM

SVM is receiving attention not only in the classification problem, and also applicable to a regression problem, accurate and applicable to various types of data of the predicted prediction easier problem. Unlike the probability estimation of the object to minimize the empirical risk, such as the logistic regression and discriminate analysis conventional methods classification SVM is got with the purpose to minimize structural risk look place only the classification efficiency itself over existing probability estimation method overall, higher predictability.

The reason of effectiveness in SVM is that it performs linear classification quickly and easily when the linearity see the non-linear data set at a low level at a higher level and expand its dimensions. Most of the pattern and cannot be linearly separated, to separate the non-linear patterns and converts the input space of the non-linear pattern in the feature space of the linear pattern. At this time, the kernel functions are used to classify non-linear patterns.

$$K(x,x_i)=((x.x_i)+1)^d \qquad (4.5.)$$

Polynomial kernel function is dependent on the direction between the two vectors, a vector having the same direction, they eventually result there is given a high value is for a polynomial kernel function, a polynomial kernel function is the same as equation (4.5).

1) *Advantages:*
a) Produce very accurate classifiers.
b) Less over fitting, robust to noise.

2) *Disadvantages*
a) SVM is a binary classifier. To do a multi-class classification, pair-wise classifications can be used (one class against all others, for all classes).
b) Computationally expensive, thus runs slow.

VI. EXPERIMENTAL DATASET

A. 20 News Groups Dataset

The 20 newsgroups dataset comprises around 18000 newsgroups posts on 20 topics split in two subsets: one for training (or development) and the other one for testing (or for performance evaluation).

1) *The 20 Newsgroups data set:* The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. To the best of my knowledge, it was originally collected by Ken Lang, probably for his Newsreaders: Learning to filter net newspaper, though he does not explicitly mention this collection. The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering.
2) *Organization:* The data is organized into 20 different newsgroups, each corresponding to a different topic. Some of the newsgroups are very closely related to each other (e.g. comp.sys.ibm.pc.hardware/ comp.sys.mac.hardware),while others are highly unrelated (e.g misc.forsale / soc.religion.christian). Here is a list of the 20 newsgroups, partitioned (more or less) according to subject matter:
3) *Data Characteristics:* One thousand Usenet articles were taken from each of the following 20 newsgroups.
   a) alt.atheism
   b) comp.graphics
   c) comp.os.ms-windows.misc
   d) comp.sys.ibm.pc.hardware
   e) comp.sys.mac.hardware
   f) comp.windows.x
   g) misc.forsale
   h) rec.autos
   i) rec.motorcycles
   j) rec.sport.baseball
   k) rec.sport.hockey
   l) sci.crypt
   m) sci.electronics
   n) sci.med
   o) sci.space
   p) soc.religion.christian

q) talk.politics.guns

r) talk.politics.mideast

s) talk.politics.misc

t) talk.religion.misc

Approximately 4% of the articles are cross posted. The articles are typical postings and thus have headers including subject lines, signature files, and quoted portions of other articles.

4) Data Format: Each newsgroup is stored in a subdirectory, with each article stored as a separate file.

B. Comparison Of Models

The performance of our categorization system is evaluated by precision, recall and F-measure.

$$P= \frac{\text{No of correctly classified documents}}{\text{No of correctly retrieved documents}}$$

$$R= \frac{\text{No of correctly classified documents}}{\text{Total No of Relevant documents}}$$

$$\text{F- MEASURE} = \frac{2*P*R}{P+R}$$

In order to compare the performance of the classification algorithms in 20 news group data set using the F-M value using a Precision, Recall and F-Measure values.

VII. CONCLUSION

Text Classification is an important application area in text mining why because classifying millions of text document manually is an expensive and time consuming task.

We analyze, that the text classification techniques such as decision trees, Bayes methods, nearest neighbor classifiers and SVM classifiers are very help full in the field of text mining, Day by day volume of electronic information is increase rapidly and extracting knowledge from these large volume data is difficult or say extracting relevant information on demand is very difficult due to large amount of data. So the main goal of text mining is to retrieve the relevant information in minimum accessing time, accurate data.

REFERENCES

[1] NingZhong, Yuefeng Li, Sheng-Tang Wu, “Effective Pattern Discovery for Text Mining,” ”, IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 1, January 2012.

[2] RashmiAgrawal, MridulaBatra, “A Detailed Study on Text Mining Techniques”, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

[3] Sonali Vijay Gaikwad, ArchanaChaugule, PramodPatil "Text Mining Methods and Techniques"International Journal of Computer Applications (0975 – 8887) Volume 85 – No 17, January 2014

[4] Fabrice Colas and PavelBrazdil , Comparison of SVM and Some Older Classification algorithms in Text Classification Tasks, Leiden University, The Netherlands, 2014

[5] Lee Junyeon, Shin Seungsoo and Kim Jungju, Comparison of Performance in Text Mining using Categorization of Unstructured Data, Indian Journal of Science and Technology, Vol 9(24), DOI: 10.17485/ijst/2016/v9i24/96148, June 2016

[6] Shuzlina Abdul-Rahman, SofianitaMutalib, NurAmiraKhanafi, AzlizaMohd Ali "Exploring Feature Selection and Support Vector Machine in Text Categorization" 16th International Conference on Computational Science and Engineering, IEEE-2013.

[7] Jiawei Han and Micheline Kamber “Data Mining Concepts And Techniques” ,Morgan kaufman publishers, San Francisco, Elsevier, 2011, pp. 285-351.

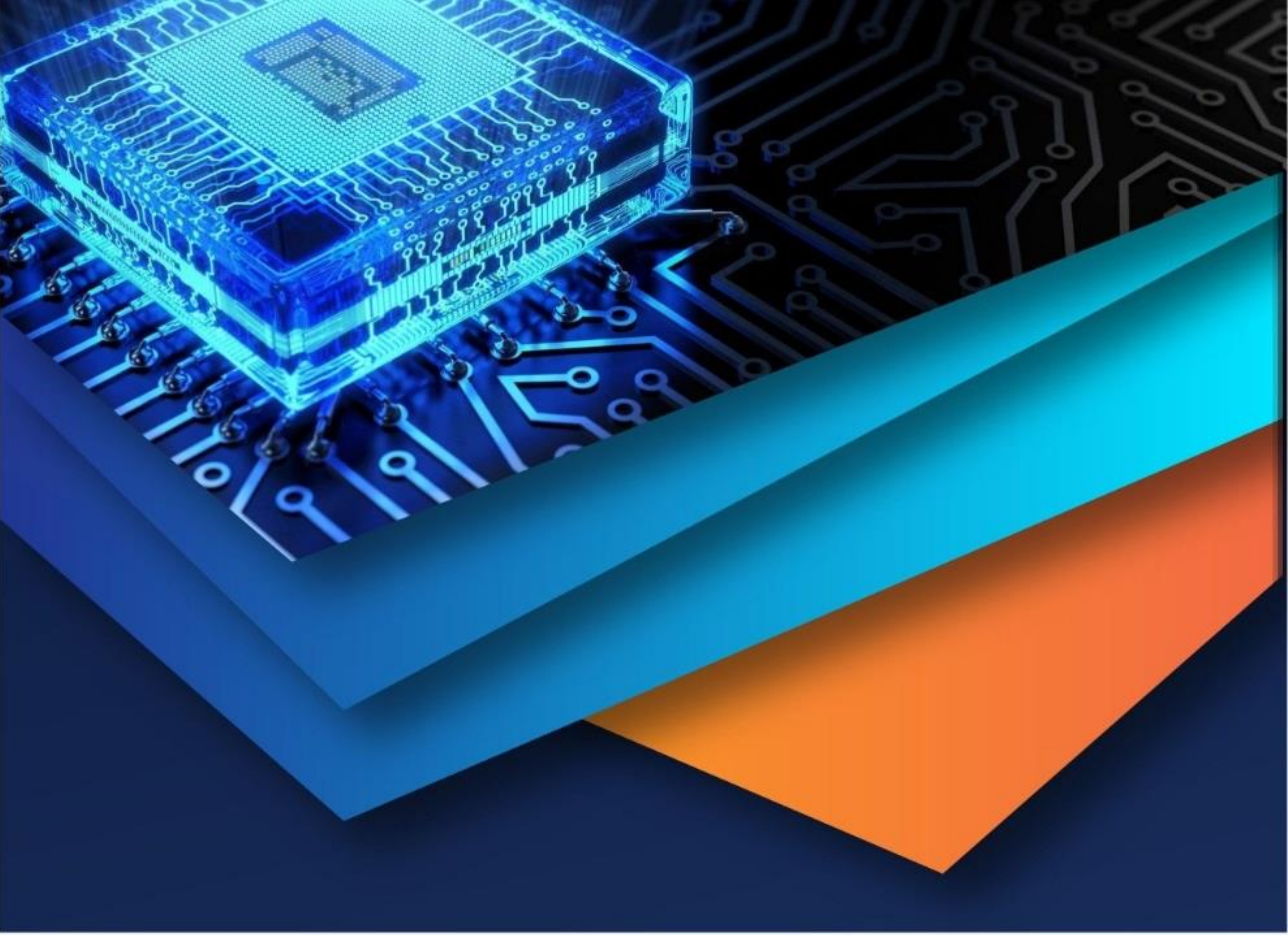
[8] Nidhi1, Vishal Gupta2"Recent Trends in Text Classification Techniques" International Journal of Computer Applications (0975 – 8887) Volume 35– No.6, December 2011.

[9] Xianfei Zhang, Bicheng Li, Xianzhu Sun "A k-Nearest Neighbor Text Classification algorithm Based on Fuzzy Integral" Sixth International Conference on Natural Computation, IEEE-2010.

BIOGRAPHY

M. Nandhiya, she received her U.G degree B.Sc(IT) from N.G.M college, Pollachi. She completed her MCA in SVS Educational Institutions, Coimbatore. Presently, she is working as a research scholar in N.G.M College, Pollachi. She presented her paper in Third International Conference.

Ms. M. Sakthi is presently working as an Associate Professor Dept of Computer Science, N.G.M College, Pollachi. She has published many papers in international/ national journals and conferences. His area of interest includes Data Mining, Parallel Computing, and Distributed Computing etc. She has 20 years of teaching and research experience.



10.22214/IJRASET

INDEX
COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

ISI

INTERNATIONAL JOURNAL
FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY