Automatic Key Term Extraction from Research Article using Hybrid Approach

Selvani Deepthi Kavila
Assistant Professor
CSE, ANITS
Visakhapatnam

B. Rajesh
Assistant Professor
CSE, GIT, Gitam University
Visakhapatnam

N. Vyshnavi
M.Tech Student
CSE, ANITS
Visakhapatnam

K. Moni Sushma Deep
M.Tech
Visakhapatnam

ABSTRACT
Key terms are subset of terms or phrases from an article that can describe the meaning of the article. In our information era, key information terms are very useful for information retrieval, article retrieval, article clustering, summarization, text mining, and text clustering and so on. These are the set of terms from an article that can describe the meaning of the article. The main aim of this paper is to help the users to quickly extract the key information automatically using hybrid systems from an article which convey the complete meaning of the text and then extracts the algorithm name present in the research paper. The focus of Hybrid system is to automatically extract the key information from various articles. Vital terms from articles are extracted by using Linguistics approaches and Statistical approaches. These terms are then passed to a rule-based extractor for further refinement where a statistical analysis is made on this set of terms according to different range of classes. Finally, this set is passed to the Multi-layered Feed Forward Artificial Neural Networks where the key information terms are extracted by using back propagation. Based on the performance evaluation, it has been observed that the acquired results are efficient when compared to manual judgement.

General Terms
Key, Information, Automatic, Retrieval, article

Keywords
Text mining, Key term extraction, Information extraction.

1. INTRODUCTION
Text Mining has become an important research area. Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. Text mining, also known as Intelligent Text Analysis, Text Data Mining, Text Analytics or Knowledge-Discovery in Text (KDT), refers to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics.

Information extraction (IE) is the task of consequently extracting organized data from unstructured and/or semi-organized machine readable articles. Automatic key information retrieval is the undertaking to distinguish a little arrangement of terms, key expressions or key terms from an archive that can depict the significance of article. Manual assignment of high quality key terms is expensive, time-consuming, and error prone. Therefore, most algorithms and systems aimed to help people perform automatic key terms extraction have been proposed.

Key information terms enable readers to decide whether an article is relevant for them or not. In our information era, key information words are very useful for information retrieval, article retrieval; article clustering, summarization, text mining and text clustering and so on. Therefore it is possible to think these terms as a set of terms or phrases semantically covering most of the text and these are the set of terms from an article that can describe the meaning of the article. Key phrases provide a brief summary of an article contents. As large article collections such as digital libraries become widespread, the value of such summary information increases. Key terms and key phrases are particularly useful because they can be interpreted individually and independently of each other. They can be used in information retrieval systems as descriptions of the articles returned by a query, as the basis for search indexes, as a way of browsing a collection, and as an article clustering technique.

As the rapid growth of textual information online, information retrieval becomes more important than ever. Key information extraction as a foundational technique for articles description has been a focal point of research in this area. Key information terms can help the users quickly refer the articles to determine whether they are worth reading or not. So key information terms are necessary. Everyday thousands of books, papers are published which makes it very difficult to go through all the text material instead there is a need of good information extraction or summarization methods which provide the actual contents of a given article. Automatic extraction of key information words doesn't need any human assistance and it speeds up the process of computation to the problems of access and discoverability, adding value to information organization and retrieval. Since key information term is the smallest unit which express meaning of entire article, many applications can take advantage of it such as automatic indexing, text summarization, information retrieval, classification, clustering, filtering, cataloging, topic detection and tracking, information visualization, report gene ratio , web searches etc.

Rest of the paper is described as follows: Section 2 describes the Literature survey based on key terms extraction, Section 3 describes the proposed system architecture and Methodology, Section 4 shows the performance analysis and results, Section 5 shows the conclusions and future work.

2. LITERATURE SURVEY
2.1 Background work related to key terms extraction
Chengzhi Z et al [1] proposed a CRF method to extract the key phrases. Cohen J. D [2] describes a strategy for drawing file terms frequency. An augmentation is likewise depicted

and showed which chooses list terms to speak to a subset of reports, recognizing them from the corpus on content is introduced. Das et al [3] proposed a neural net model used to pre-process a data string and match with the client characterized string. They extracted highlighted terms from the given content with the client characterized terms .If there is a match then the estimation of that term increments. This procedure rehashes until it accomplishes a steady esteem and aggregate sentence score. Damien H et al [4] proposed a text mining strategy that depends on key terms and key sentences area to remove the primary data required by the reader and afterward stores the separated data in database for further handling. Ercan G, Cicekli I [5] describes the issue of programmed extraction of key phrases from archives is dealt with as a directed learning task. A lexical chain holds an arrangement of semantically related expressions of content and it can be said that a lexical chain speaks to the semantic substance of a part of the content. Frank Eet al [6] demonstrates that a straightforward system for key phrase extraction in view of the naive Bayes learning plan performs similarly to the cutting edge. Ion Muslea[7] depicts an overview on different sorts of extraction examples that are created by machine learning algorithms. Jasmeen Kaur, Vishal Gupta [8] depicts the strategies have been exhibited that can be connected to remove successful key phrases that extraordinarily distinguish an article. Kamal Sarkar, Mita Nasipuri and Suranjan Ghose [9] portray a neural system based way to deal with key phrase extraction from investigative articles. Menaka S and Radhika N [10] proposed a Text Classification using Keyword Extraction Technique and Content grouping is one of the real utilizations of machine learning. They proposed a technique use content mining calculations to concentrate key terms from research papers. Mihalcea R an Tarau P[11] proposed that accuracy can be attained by considering methods like key term extraction and sentence extraction. Graph-based algorithm is designed and it is successfully applied in analysis. Naidu Reddy [12] proposed a summarization technique, are used to extract the crucial information and provide a synopsis for the news in the e-Newspaper. Parmar paresh B. and Ketan patel [14] explains that Quality and speed are the key factors on which a processing application depends on in a key term extraction algorithm. It uses key term phrases to select the important sentence. Rahul B Diwate, Satish J. Alaspurkar [15] proposes an investigation and correlation of distinctive algorithms for full pursuit identical example coordinating like unpredictability, effectiveness and systems. Raymond et al[16] proposed a Discovery from Text Extraction approach, for separating an organized database from a content corpus. This methodology utilizes a programmed learned Information Extraction framework and after that mines this database with existing KDD apparatuses. Yang Shansong [17] Key Phrases can be used to organize the relevant articles and it's explanatory to derive correlations among articles which belong to similar aspects of research.

2.2 Existing techniques

There are two existing ways to deal with automatic key term indexing [13]

2.2.1 Key term extraction

Terms occurred in articles are analyzed to identify the significant term such as length and frequency. Here point is to extract data as for their importance in text without earlier vocabulary.

2.2.2 Key term Assignment

Key terms are chosen from a controlled vocabulary of terms and articles are grouped according to their content into classes that relate to components of vocabulary. This methodology is called Text Categorization. There is an earlier arrangement of vocabulary and point is to match them to messages in a set. Existing methods for Automatic Key terms Extraction can be categorized into four:

1. Simple Statistical Approach
2. Linguistics Approach
3. Machine Learning Approach
4. Other Approaches

3. PROPOSED TECHNIQUE

The proposed technique used in this paper is Hybrid technique. It is the combination of Statistics, Linguistics, Rule - Based and Back propagation algorithm. It employs more than one technique to solve a problem as shown in Fig.1. The following algorithms are used for implementation and research papers are used as data set.

3.1 Linguistics Approach

This approach uses the phonetics such as highlighting of the terms, sentences etc. It is descriptive rather than prescriptive. Analysis of a language is included in a linguistic approach and particularly it determines the nature of language. Text processing can be manually done with-out the use of an algorithm. Present knowledge is evaluated using certain techniques and processes. It includes the syntactic analysis, discourse analysis, lexical analysis, etc. In this work, linguistics approach applied is the parts of speech tagging (POS) where the terms in the entire article is tagged with parts of speech and finally only noun phrases are taken.

3.1.1 Noun Phrase Extraction

Noun phrases are the combination of nouns, adjectives and both the combination of the nouns and the adjectives. The steps followed in implementing the noun phrases are that, firstly the various conditions for tagging the nouns, adjectives, adverbs and verbs is implemented. Finally the combination of the nouns, adjectives and both nouns and adjectives are taken as noun phrases.

The proposed algorithm is presented below in Table 1.

Table 1 Steps for linguistic approach

Algorithm
Step 1: read the array list
Step 2: read the nouns list
Step 3: tag the terms which are nouns as \\NN
Step 4: read the adjectives list
Step 5: tag the terms which are adjectives as \\ADJ
Step 6: read the adverbs list
Step 7: tag the terms which are verbs as \\VB
Step 8: read the verbs list
Step 9: tag the terms which are adverbs as \\ADV
Step 10: read the adverbs list
Step 11: read the terms which are nouns, adjectives and adjective + noun
Step 12: tag the terms as \\NP
Step 13: display the noun phrases which are tagged as \\NP

3.2 Statistical Approach

The aim of the statistical analysis is to calculate the probability that differences as great as or greater than those observed could be due to chance. These methods are simple and do not need the training data. The statistics information of the terms can be used to identify key information term in the article. The statistics methods include N-gram statistical information, term co-occurrences, TF*IDF, term frequency and PAT- tree, etc.

Table 2 Steps for Statistical approach

Algorithm

Step 1: read the arraylist
Step 2: calculate the length of the term
Compute Len ← Len(term)/(Max_Len)
Step 3: calculate the position of the term
Compute term = #(term)/(∑ $term_i$)
Step 4: calculate the phrase frequency of the term
Compute F(freq)= sqrt(0.5*pf*pf*plc)
Step 5: find the probability for t and a
Step 6: check whether the term is present in the title or not
if present then t = 1, otherwise t = 0
Step 7: check whether the term is present in the abstract or not
if present then a = 1, otherwise a = 0
Step 8: print the features list

3.3 Rule – Based approach

The Rule – based approach is a method where the learned model is represented as a set of IF –THEN rules. Rules are a good way of representing the information or bits of knowledge. In this work, the rule – based approach is employed in order to establish the knowledge base and key information terms has been divided by depending on the phrase frequency, abstract and title features.

Table 3 Steps for Rule – Based Approach

Algorithm

Step 1: read the array list

Step 2: find whether the frequency of the term lies in the range as per the user requirement

Step 3: the find if the term is present in both the title and abstract or not

Step 4: if so then pass the target value as 1

Step 5: otherwise pass the target value as 0

Step 6: display the result according to the user requirement

3.4 Back Propagation Algorithm

The Back propagation is the neural network learning algorithm. A neural network is a set of connected input/output units in which each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to able to predict the correct class label of the input tuples. The back propagation algorithm performs learning on a multi – layer feed forward neural network. It iteratively learns a set of weights for the prediction of the class label of tuples. A multi layer feed – forward neural network consists of an input layer, one or more hidden layers and an output layer. In this paper, five input nodes, five hidden nodes and a single output node are to be taken. The target for the model is fixed according to the knowledge base established according to the rule – based extraction and the statistical analysis applied. This is implemented as follows: the weighted words with numerical weights that are found out form the features will be given as the input for the feed – forward network, then learning process is followed and the error is calculated. In order to calculate the error the above target value is used (either 1 or 0). Then the error rate is compared with the threshold value that has been fixed. If the error rate lies below the threshold the process is terminated and the output is given otherwise the weight are updated accordingly and the process is back propagated again and again the error rate is calculated and this process is iteratively followed till the error rate falls below the threshold value. The output of this model is the key information words.

Back propagation algorithm is a n-n-1 type network. It represents that input layer would contain 'n' nodes, which will be equal to the number of features that are applied. The number of hidden nodes is equal to the number of the input nodes and finally a single output node which resembles whether a particular term is key information term or not.

Table 4 Steps for Back propagation

Algorithm

Step 1: read the features list

Step 2: Initially the parameters of the Back propagation algorithm i.e. hidden layer nodes, output layer nodes, learning rate and angular momentum are accepted.

Step 3: Weights for the connections of the network are accepted randomly between ranges of -1 to 1.

Step 4: The number of epochs, the model has to run is accepted.

Step 5: The dataset is partitioned into training set and testing set by using hold out method

In the Hold out method the given data are randomly partitioned into two independent sets, a training set and a testing set. Typically, two – thirds of the data are allocated to the training set and the remaining one – third is allocated to the test set. The training set is used to derive the model, whose accuracy is estimated with the test set. The estimate is pessimistic because only a portion of the initial data is used to derive the model.

Step 6: The training data are sent for training and the weights are updated and these weights are used for calculation of output values of testing data.

Step 7: Finally the required key information terms are extracted

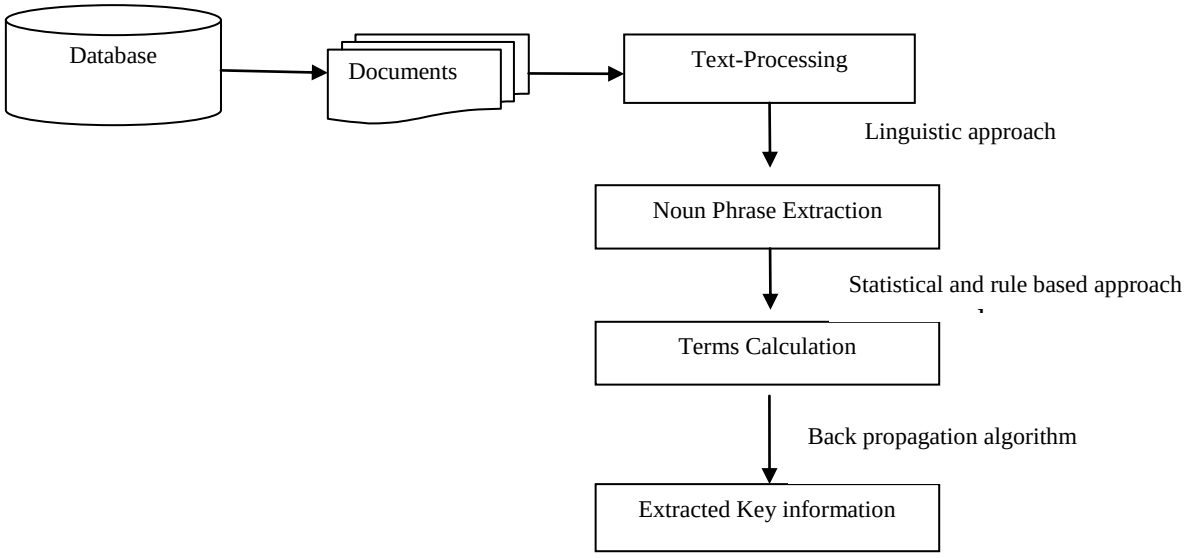

Fig.1 System Architecture

4. PERFORMANCE ANALYSIS AND RESULTS

It is essential to measure the performance for information retrieval system that has been utilized as parts of the paper are described. To assess the results, a few standard measures are to be considered such as Accuracy and Error rate. Accuracy is the condition of being true it is a measure of correctness. It is the correct value of a standard which is measured. Error rate is calculated from the true values and measured values. Error rate must be reduced with increase in the count of papers. Similarly, Accuracy gets increased with increase in papers. These two entities are considered as metrics for improving performance in case of information extraction.

$$\%Error = \frac{\text{True value-Measured value}}{\text{True value}} * 100$$

$$\%Accuracy = (100\%) - (\%Error)$$

In the repository, few articles are maintained and for these articles the key terms are extracted. Here manual judgment results are compared with the system generated results.

A manual judgment is generated for all the research papers .A sample judgment file is shown below for the first 10 papers that are studied.

Table 5 Comparison of manual and system generated judgment

Document no	Manually generated key terms	System generated key terms
D1	5	4
D2	5	4
D3	3	3
D4	5	6
D5	7	5
D6	5	4
D7	4	5
D8	4	3
D9	5	4
D10	5	4

The values Accuracy and Error rate are calculated as shown in table 6,

Table 6: Calculation of Accuracy and Error rate for articles

No. of articles	Accuracy	Error
10	87.5%	12.5%
20	87.9%	12.3%
30	89.2%	11.4%

The table 6 depicts the performance of the system for all the different classes of papers for key term extraction.

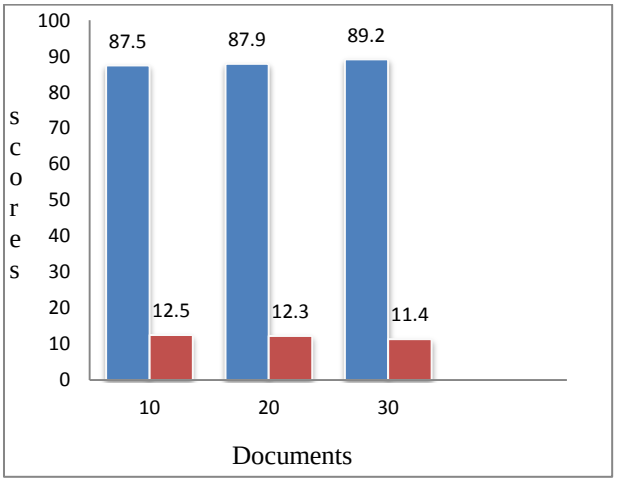

Fig. 2 Accuracy and Error of the system

Fig. 2 depicts the accuracy and error of the system. This shows that the results obtained are significantly better. The x-axis represents the count of 'N' articles and y-axis represents the score of accuracy and error. The system generated nearly 87% percent accuracy.

5. CONCLUSION AND FUTURE WORK

This work mainly focuses on the automatic key information extraction from the articles using the Hybrid system which employs various techniques like Linguistics Approaches, Statistical Approaches, Rule – Based Approach and back propagation algorithms. By employing the Hybrid approach it is possible to prevent the limitations and drawbacks of every individual technique. The conclusion of this work enables the user to extract the important information from the required article such as key terms.

The extension of the work can be done by implementing the combination of the soft computing techniques and the Neural Networks approach in order to get more accuracy in the information extraction and by incorporating some more important features.

6. REFERENCES

[1] Chengzhi Z ,Huilin W et al, "Automatic Keyword Extraction from Documents Using Conditional Random Fields", Journal of Computational Information Systems, Volume 4, issue 3, (2008).

[2] Cohen J.D, "Highlights: Language and Domain-independent Automatic Indexing Terms for Abstracting" Journal of the American Society for Information Science, Volume 46 issue 3, pg no: 162-174, (1995).

[3] Das A, Marko M et al, "Neural Net Model For Featured Word Extraction", Neural and Evolutionary Computing, ACM, (2002).

[4] Damien Hanyurwimfura, Bo Liao et al, "An automated Cue Word based Text Extraction" Journal of Convergence Information Technology (JCIT), Volume7, Number10,(2012).

[5] Ercan G, Cicekli I, "Using Lexical Chains for Keyword Extraction", Information Processing and Management, Volume 43 Issue: 6, pg no: 1705-1714, (2007).

[6] Frank E, Paynter G.W, Witten I.H, "Domain-Specific Key phrase Extraction" Proceedings of the 16th International Joint Conference on Artificial Intelligence, Sweden, pg.no: 668-673, (1999).

[7] Ion Muslea, "Extraction Patterns for Information Extraction Tasks: A Survey", AAAI Technical Report WS-99-11.

[8] Jasmeen Kaur, Vishal Gupta, "Effective Approaches For Extraction of Keywords", IJCSI International Journal of Computer Science, Volume 7, Issue 6, (2010).

[9] Kamal Sarkar, Mita Nasipuri and Suranjan Ghose, "A New Approach to Key phrase Extraction Using Neural Networks", IJCSI International Journal of Computer Science Issues, Volume 7, Issue 2 No 3, (2010)

[10] Menaka S, Radha N, "Text Classification using Keyword Extraction Technique", International Journal of Advanced Research in Computer Science and Engineering, Volume 3, Issue 12, (2013).

[11] Mihalcea R and Tarau P, "Text rank: Bringing order into texts", Association for computational linguistics, (2004).

[12] Naidu Reddy et, al "Text summarization with automatic key word extraction in Telugu E-News Papers", (2017).

[13] O. Medelyan, I. H Witten, "Thesaurus Based Automatic Key phrase Indexing", in Proceedings of the Joint Conference on Digital Libraries 2006, pg.no-296-297, Chapel Hill, NC, USA, (2006).

[14] Parmar Paresh B and Ketan Patel "A Survey Paper on Mining Keywords Using Text Summarization Extraction System for Summary Generation over Multiple Documents" Volume 5 Issue 11, (2016).

[15] Rahul B. Diwate, Prof. Satish J, Alaspurkar, "Study of Different Algorithms for Pattern Matching", International Journal of Computer Science (IJCSI) Volume 7, Issue 2, No 3, (2010).

[16] Raymond J. Mooney and Un Yong , "Text Mining with Information Extraction" Proceedings of the 4th International MIDP Colloquium, (2003).

[17] Yang, Shansong et. al "Key phrase DS: Automatic generation of survey by exploiting key phrase information", Volume 224, (2017).