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. 4, Issue 11, November 2016

Keyword Mining and Co-Word Construction by Proposed Tagging, Statistical and Graph Based (TSG) Algorithm

R. Nagarajan[1], S. Anu H Nair[2], P. Aruna[3], N. Puviarasan[4]

Assistant Professor/Programmer, Dept. of Computer Science & Engineering, Annamalai University, Tamilnadu, India[1]

Assistant Professor, Dept. of Computer Science & Engineering, Annamalai University, Tamilnadu, India[2]

Professor, Dept. of Computer Science & Engineering, Annamalai University, Tamilnadu, India[3]

Associate Professor, Dept. of Computer Science & Engineering, Annamalai University, Tamilnadu, India[4]

ABSTRACT: Keyword mining, a challenging job in managing unorganized documents. Nowadays, tremendous amount of enrich information is available dynamically in the form of unstructured documents, organizing such a documents is the major challenge. Clustering the unstructured documents based on their concepts is the best and useful way to organize documents. Concept extraction plays the vital role in document clustering. Keyword mining is one of the best methods for concept extraction. In this research paper, TSG (Tagging, Statistical and Graph based) algorithm is proposed for keyword mining. This proposed algorithm integrates the influence of statistical and graphical techniques with tagging approach. This proposed algorithm awards 90% of accuracy.

KEYWORDS: Data Mining, Document Clustering, HTML Tagging, Statistical methods, Graph based Methods, Keyword extraction, Co-words

I. INTRODUCTION

Keywords play a vital role in document clustering. This research work proposed a novel keyword mining algorithm. In this proposed algorithm, keyword mining is achieved by calculating the term weight not only based on their frequency, but also on their physical position of document such us major title, subtitle, italic quotations, bold specification, author keywords etc. It is also found through our research that mere a single keyword is not enough for document clustering. Because of this reason, our second part of proposed algorithm constructs the co-words from the weighed keywords. Our research work relies upon the assumption that (i) authors of scientific articles choose their technical terms carefully; (ii) when different terms are used in the same articles it is therefore because the author is either recognizing or postulating some non-trivial relationship between their references; and (iii) different authors appear to recognize the same relationship, then that relationship may be assumed to have some significance within the area of science concerned. In this proposed system, influence of HTML tags is used to extract the keywords effectively.

II. RELATED WORK

This section reviews the previous work on keyword mining algorithms and discusses the difference between them. In [1] proposed a method for Keyword based Tweet Extraction and Detection, this method collects tweet using a specific keyword and then summarize them to find out topics related to that keywords. The Topic detection is done by using clusters of frequent patterns. In [2] proposed Conditional Random Field based keyword extraction approach. CRF model is a state-of-the-art sequence labelling method, which uses the features of documents more sufficiently and effectively and at the same time keyword extraction can be considered as the string labelling. In [3] instead of usual tf.idf score for keywords, some alternative relevance measures that do use relations between words has been used and keyword weights are computed by defining co-occurrence distributions for words and comparing these distributions with the document and the corpus distribution. In [4] proposed a graph based ranking algorithm called as Semantic

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. 4, Issue 11, November 2016

Rank for keyword and sentence extraction from text. This proposed algorithm constructs a semantic graph using implicit links, based on semantic relatedness between nodes and consequently ranks nodes using different ranking algorithms. In [5],) suggested guidelines for selecting and processing keyword sets for using text-mining in patent analysis. Four different factors have been considered as, Which element of the patent documents to adopt for keyword selection, What keyword selection methods to use, How many keywords to select and How to transform the keyword selection results into an analyzable data format, to identify the best strategy, experiment based on an orthogonal array of the four factors was designed, in which the four factors were evaluated and compared through *k*-means clustering and entropy values. In [6] different methods for keyword extraction and the impact of the number of keywords on the quality of the incremental clustering have been discussed. In [7] presented and validate two graph-based methods, supervised and unsupervised graph based syntactic representation approaches, for cross-lingual keyword extraction to be used in extractive summarization of text documents. In [8], step by step text clustering framework and challenges faced by text clustering methods has been discussed. In [9], varies similarity measures and distance measuring function has been analyzed, result comparison with standard k-means algorithm has also been discussed. In [10], proposed a constrained PLSA algorithm for tagged documents with small amount for seed and features were represented in four combinations of tags and abstraction of words.

III. Proposed TSG (Combination of Tagging, Statistical, Graph) Keyword Mining Approach

The proposed TSG keyword mining algorithm has two parts; First part of the proposed algorithm concentrates on the extraction of keywords from the documents. The second part of the algorithm constructs the co-words from derived mined keywords.

In the first part, initially documents (D_i) taken for clustering are converted into HTML documents (HD_i) using standard HTML converter. The next step is tokenization. Terms of the documents are extracted along with their HTML tags. By referring the STEMMAST database, which contains the fruitful stems which are derived by the standard porter stemming algorithm, stemming process of the derived terms are conducted. Similarly, STOPMAST reference database which contains the standard stop words is used to remove the stop words which carry no information. Followed by stemming, stop word removal process is conducted on the resulted documents, during the stop word removal process, the words of the documents are referred with STOPMAST database which consists of the standard stop words collected from the various document analysis. Fig. 1 shows the stop word removal process of the sample document.

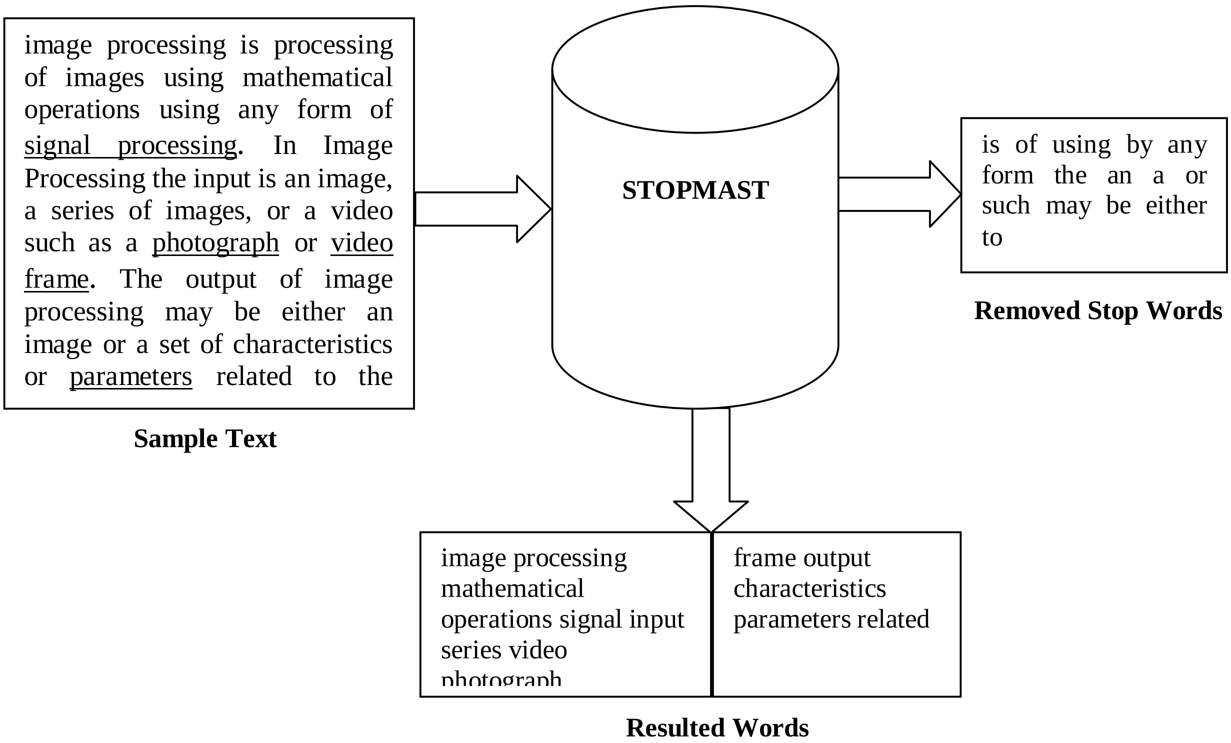

Fig. 1 Stop word removal process

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. 4, Issue 11, November 2016

Then, the resulted terms are identified under three groups using HTML tags as follows,

- Title Group Terms
- Property Group Terms
- Raw Group Terms

Title Group Terms (TG)

Title group terms are the terms physically used in the title, subtitle, table, figures, images, author keywords of the documents. These title group terms are extracted with their frequencies and document ID by the HTML Tags such as <Title>, <head1> , <head2>, , <head3>, <Img> etc.

Property Group Terms (PG)

Property group terms shows the properties of the term such that bold, bolditalic, italic, colored, big, strong, subscripts, superscripts, small, underlined, strikethrough, etc., These terms are extracted with their frequencies, document ID and sentence ID by using the HTML tags like <B>, <I>, <BI>, <FONTCOLOR>, <EM>, <BIG>, <STRONG>, <SMALL>, <Caption>, <STRICKE> etc.,

Raw Group Terms (RG)

The remaining terms, usually used in the body of the document are raw group terms. Raw group terms are extracted with their frequencies, document ID and sentence ID using the HTML tags like <text>, etc. Fig. 4.3 shows the flow of proposed TSG keyword extraction method.

Table 1: HTML Tags used to extract Title Group (TG) Terms

1	<Title>	</Title>	Defines the document title
2	<head>	</head>	Defines information about the document
3	<head1>, <head2>…	<head1>, <head2>…	Defines Sub Titles
4	<span>,<div>	<span><div>	Defines header 3
5	<abbr>	</abbr>	Defines an abbreviation
6	<acronym>	</acronym>	Defines an acronym
7	<blockquote>	</blockquote>	Defines a long quotation

Table 1 illustrates some commonly used HTML Tags for Title Group Terms extraction used in this research work.

Table 2: HTML Tags used to extract Property Group (PG) terms

1	<B>	</B>	Defines bold text
2	<BI>	</BI>	Defines bold italic text
3	<I>	</I>	Defines italic text
4	<Font>	</Font>	Defines text font, size, and colour
5			Defines strong text
6	<Big>	</Big>	Defines Big text
7	<Small>	</small>	Defines Small text
8	_		Defines subscripted text
9	<strike>	</strike>	Deprecated. Defines strikethrough text
10	<caption>	</caption>	Defines a table caption
11	<Img>	</Img>	Defines an image
12			Defines emphasized text

Table 2 shows the property group HTML Tags used in this research work to extract terms based on their properties.

Table 3: HTML Tags used to extract Raw Group (RG) terms

1	<text>	</text>	Defines a text area
2	<body>	</body>	Defines the body element

Table 3 lists few commonly used HTML tags to extract raw text from the 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. 4, Issue 11, November 2016

Algorithm 1: Proposed TSG approach

Input	Document corpus D_i
Output	A ranked keywords of the documents under three groups (TG,PG & RG)
Step 1	Documents (D_i) are converted into HTML Documents
Step 2	Tokenization of the documents (stemming & stop word removal (by STEMMAST & STOPMAST)
Step 3	Terms are extracted under three groups
Step 4	Term in title, subtitle, tables, figures, images etc., are extracted with their document ID and labelled as title group terms.
Step 5	Terms in bold, italic, bold italic, coloured etc., are extracted with their document ID and sentence ID and labelled as property group terms.
Step 6	Terms in body of the documents are extracted with document ID and sentence ID and labelled as raw group terms.

Step 7	****Applying Statistical Approaches-High Frequency & Term Frequency- Inverse Sentence Frequency to extract keywords****
	***** TG – Title Group Terms, PG – Property Group Terms, RG – Raw Group Terms*
	For all $T_j \in TG, PG \;\&\; RG$
	Begin,
	    if Title group terms
	        Rank-TG = Rank(HighFreq($T_j TG$))
	    endif
	    if Property Group Terms
	        Rank-PG=Rank(HighFreq(T_jPG) & TF-ISF(T_jPG))
	    endif
	    if Raw Group Terms
	        Rank-RG=Rank(HighFreq(T_jRG) & TF-ISF(T_jRG))
	    endif
	End for
	End

Rank()

Step 1 Get the high frequency Terms from HighFreq()
Step 2 Ranks higher frequency terms
Step 3 Return Ranks with Frequency

HighFreq()

Step 1 Receives the set of terms with their frequencies
Step 2 if Frequency of the term >= Threshold Value
Step 3 Hfreq-terms = Hfreq-terms + term
Step 4 endif
Step 5 Return Hfreq-terms

TF-ISF()

Step 1 Receives the set of terms with their frequencies and Sentence ID
Step 2 Calculate TF(Terms) ; Calculate ISF(Terms) ; Calculate TF-ISF(Terms)
Step 5 if TS-ISF(Terms) >= threshold Value
Step 6 TFISF-Term = TFISF-Terms + Term
Step 7 endif
Step 8 Return TFISF-Terms

At this stage the proposed algorithm ranked the list of title group terms, property group terms and raw group terms with their frequencies, sentence ID and document ID. The Next step of our proposed work is to apply statistical approaches to these three groups of terms.

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. 4, Issue 11, November 2016

The statistical approaches used here are,

1. High Frequency (HF)
2. Term Frequency Inverse Sentence Frequency (TF-ISF) and
3. Co-occurrence Statistical Information (CSI).

The First two approaches are used to identify the important keywords from the documents. The third method CSI is used to construct co-word. Co-words are the combination of two or more important keywords, which contribute more to find out the concept of the documents.

i) High frequency statistical approach on title group terms, property group terms & raw group terms

Usually title group terms are extracted from the title, sub title, heading, caption of the table and figures etc., of the documents. Hence only high frequency statistical approach is considered. Then, the proposed algorithm ranked the top ranking title group terms within the threshold value, is identified as Keywords of the Title Group Terms. Likewise top ranked property group terms are ranked based on their frequency within the threshold value as keywords of the property group terms and top ranked raw group terms are ranked as keywords of the raw group terms. Now, the proposed algorithm derives the top ranked keywords of all title, property and raw group categories with their ranks.

ii) Term Frequency-Inverse Sentence Frequency statistical approach on property group terms and raw group terms

TF-ISF statistical approach is applied on property group terms and raw group terms only, because title group terms are never presented in sentences. This method is inverse function method. Because of this inverse nature, this approach eliminates the words, even though having more frequencies but not much contributed to find out the concept of the document. For example, word appears more number of times in a single sentence is not an important keyword. By comparing the top ranked high frequency words with the top ranked TS-ISF words, the common words of both property group and raw group terms are identified and labelled as property group keywords and raw group keyword. After applying the statistical approaches to our three group terms, first part of our proposed algorithm derives the ranked keyword set for title group, property group and raw group. The second part of the proposed algorithm constructs the co-words from the derived keywords. Co-words are constructed by identifying the binding or linkages or strength among the derived keywords. The proposed algorithm finds the linkage strength between the words by three different approaches. First approach is co-occurrence statistical information (CSI), which is statistical approach. Second method is graph based co-word construction approach and the last approach is co-occurrence matrix approach. The following Algorithm 3 explains the co-word construction process using co-occurrence statistical information (CSI) and graph based method.

Algorithm 2: Proposed co-word construction algorithm (CSI & Graph based approaches)

Input	Ranked keyword sets of title group (TG), property group (PG) and raw group (RG)
Output	co-words

Step 1	** *Co-occurrence statistical Information and graph based approach for co-word construction* ***
	Begin
	For $all\ T_j \in Rank - TG$ // T_j is the jth term, T_k is j+1th term
	Coword-TG = $CSI(Rank - TG(T_jT_k))$
	Endfor
	For $all\ T_j \in Rank - PG$
	Coword-PG = $CSI(Rank - PG(T_jT_k))$
	Endfor
	For $all\ T_j \in Rank - RG$
	Coword-RG = $CSI(Rank - RG(T_jT_k))$
	Endfor
	CSI_Co-word = Co-word-rank(Coword-TG, Coword-PG, Coword-RG)
	End
	Output **CSI(Co-words)**

Step 2	Co-Word_Keyword_Graph_based = Construct_graph(Coword_TG,Coword_PG,Coword_RG) Output **Graph(Co-Words)**

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. 4, Issue 11, November 2016

Step 3	Compare the Top Ranked Keyword/Co-word of both CSI & GRAPH results
Step 4	Top Common Keyword/Co-Words are extracted and labelled as KEYWORDS of the documents
	Return **Co-Words** with their Linkage Strength

```
Co-word-rank()
        Step 1  Get the co-words
        Step 2  Ranks higher frequency terms
        Step 3  Return co-words with their ranks & frequency

Construct_graph()
        Receive Ranked_keywords ( Rank_TG, Rank_PG, Rank_RG)
        Initially G an empty Graph

        For all Keyword_co-word (Vi)
            If empty of Graph G
                Initialize G with Vi
            Else
                If Ranked_keywords(Vi) not exist in G
                    Add Vi to G
                endif
            endif
        End for

        For all pairs of vertices (Vi,Vi)
            Calculate Linkage Strength(Vi,Vi)
            Draw Edges between Vi & Vi  (based on linkage strength)
            Construct co-words based on edges
End for
```

The Second part of our proposed algorithm also takes ranked keyword sets of title group, property group and raw group as input. The co-occurrence statistical information (CSI) is applied on the entire three group ranked keywords separately. By this CSI application, the linkage strength between the words is calculated, and the higher CSI value words constructs co-words. The co-words constructed by CSI in all the three group words are compared and the top value co-words are identified as keyword/co-words of the documents and labelled as CSI based Keyword.

The next step of the algorithm constructs the co-words based on graph based approach. This is because of to get the quality co-words from documents. In this graph based approach, undirected graph is constructed for every document; the keywords of the documents are represented as nodes of the graph and co-occurrence relation between the words are represented as edges. Syntactic filters are used to filter the nodes of the graph. Edge weights are calculated as the co-occurrence count of the words they connected. At the end, the proposed algorithm constructs the co-words of the documents by analyzing the graph constructed. This keywords/co-words are labelled as graph based keywords.

By comparing the list of keywords/co-words derived from CSI, graph based and co-occurrence matrix approaches top ranked common co-words and single words are identified as the keywords of the document. These keyword set contains both keyword and co-words.

IV. EXPERIMENTAL RESULTS

A. *KEYWORD MINING:*

To validate our proposed TSG algorithm, sample of 60 documents with various filed of computer science were taken. Experiments were conducted in two phases. First phase of the proposed algorithm mined the sample documents and extract the features from them. The second phase constructs the co-words from the mined keywords. As per the proposed algorithm, initially all the 60 documents were converted into HTML documents, using the specific HTML tags. Words of the documents were extracted. Using STOPMAST stop words were removed and similarly words were tuned by using STEMMAST. After that, total 3202 unique words were extracted with their frequency and sentence Id,

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. 4, Issue 11, November 2016

among those 121 words were identified as Title Group (TG), 172 words were identified as Property Group (PG) and 2909 words were identified as Raw Group (RG). Ranked Title Group words, Property Group words and Raw Group words were calculated by applying the statistical approaches High Frequency and Term Frequency-Inverse Sentence Frequency on minded words.

Table 4 : Sample Documents and its words extracted by CSI approach

S.No.	Data Set	Size	Unique Words	Title Group	Property Group	Raw Group
1	Image Processing	12	684	31	39	614
2	Data Mining	13	797	35	41	721
3	Computer Graphics	12	598	22	35	541
4	Computer Networks	11	551	16	27	508
5	Software Engineering	12	572	17	30	525

Table 4, describes the data set taken for experiments. Different fields of computer science such as Image processing, Data mining, Computer Graphics, Computer Networks and Software Engineering are considered for experiments. In image processing documents, after applying the proposed algorithm, 684 unique words are extracted, in which those 31 numbers of words are identified as title group words, 39 as property group words and 614 words as raw group words. Likewise, the remaining domain documents are experimented with the proposed algorithm.

B. *CONSTRUCTION OF CO-WORDS:*

The Second phase of our proposed algorithm takes ranked keyword sets of title group, property group and raw group as input. The co-occurrence statistical information (CSI) was applied on the entire three group ranked keywords separately. By this CSI application, the linkage strength between the words was calculated, and the higher CSI value words constructs co-words. Totally, 228 co-words were constructed by comparing the top-value co-words of all the three group Title group, Property group and Raw group. These 228 co-words were labelled as CSI based Co-words.

To get more accuracy, graph based approach was applied on these CSI based co-words. In graph construction, CSI co-words were taken as input, the linkage strength between the co-words were calculated using the equation 1. The strength of the linkage ‘S’ of the co-occurrences are measured by using the

$$S(w_i w_{ij}) = w_{ij} * w_{ij} / w_i w_j \qquad (1)$$

where W_{ij} is the number of co-occurrence of words i and j.

The higher (1) or the lower (0) strength of the co-occurrence is determined by the number of times a co-word pair occurs together in a given document. As the strength of the co-occurrence ranges between 0 and 1, it is necessary to fix certain thresholds in order to filter the 'uninteresting co-occurrences'. This implies that the co-word pairs with a co-occurrence frequency under a certain threshold value are filtered out. The higher the occurrence of the co-word pair, the greater will be the strength of their association and vice-versa. Here the value 5 is fixed as threshold value. From the strength of association one could identify the semantic relationship existing between the terms. In this manner, co-words for the entire set of sample documents were collected with their parent document identification. Edges were drawn based on the linkage strength and finally 178 co-words were identified as the co-words of the sample documents.

V. CONCLUSION

In this paper, a procedure for keyword mining by extracting title group (TG), property group (PG) and raw group (RG) terms have been established. The algorithm for the keyword mining and construction of co-word has been proposed. A new combined Tagging, Statistical and Graph based (TSG) keyword mining approach has been introduced using the proposed constructed co-word.

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. 4, Issue 11, November 2016

REFERENCES

1. Amrutha Benny and Mintu Philip, “Keyword Based Tweet Extraction and Detection of Related Topics”, Procedia Computer Science (Elsevier), Vol.46, pp.364-371, 2015.
2. Chengzhi Zhang, Huilin Wang, Yau Liu, Dan Wu, Yi Liao and Bo Wang, “Automatic Keyword Extraction from Documents using Conditional Random Fields’, Journal of Computational Information Systems, Vol. 4, Issue 3, pp. 1169-1880, 2008.
3. Christian Wartena, Rogier Brussee and Wout Slakhorst, “Keyword Extraction using Word Co-occurrence”, Workshops on Database and Expert Systems Applications (IEEE), pp-54-58, 2010.
4. George Tsatsaronis, Iraklis Varlamis and Kjetil Norvag, “SemanticRank: Ranking Keywords and Sentences using Semantic Graphs”, 23rd International Conference on Computational Linguistics-Coling 2010, pp. 1074-1082, 2010.
5. Heeyong Noh, Yeongran Jo and Sungjoo Lee, “Keyword selection and processing strategy for applying text mining to patent analysis”, Expert Systems with Applications (Elsevier), Vol. 42, Issue 9, pp. 4348-4360, 2015.
6. Rafael Geraldili Rossi, Ricardo Marcondes Marcacini and Solange Oliveira Rezende, “Analysis of Statistical Keyword Extraction Methods for Incremental Clustering”, Grant 2011/12823-6 and 2011/19850-9 (FAPESP), 2011.
7. Marina Litvak and Mark Last, “Graph-Based Keyword Extraction for Single-Document Summarization”, Coling 2008: Proceedings of the workshop Multi-source Multilingual Information Extraction and Summarization, pp. 17-24, 2008.
8. Francis M. Kwale, “An Efficient Text Clustering Framework”, International Journal of Computer Applications, Vol. 79 No. 8, 2013.
9. Anna Huang, “Similarity Measures for Text Document Clustering”, New Zealand Computer Science Research Student Conference, pp. 49-56, 2008.
10. Chien-Liang Liu, Wen-Hour Hsaio, Chia-hoang Lee and Chun-Hsien Chen, “Clustering tagged documents with labeled and unlabeled documents”, Information Processing and Management, Vol. 49, pp. 596-606, 2013.

BIOGRAPHY

R. Nagarajan was born in 1970 in India. He received his M.C.A and M.Phil. (Computer Science) from Bharathidasan University, Tiruchirapalli. Presently he is working as Assistant Professor/Programmer in Annamalai University. His area of research is Data Mining and Document Clustering. He has fifteen years of experience in Application Programming. He is involved in research activities for the past Nine years.

Dr. S. Anu H Nair is an Assistant Professor in the Department of Computer Science and Engineering of Annamalai University. She received her Bachelor degree in 2002, Master Degree from Manonmaniam Sundaranar University in 2005 and the Ph.D degree from Annamalai University in 2016. Her main research areas include image processing, pattern classification techniques and neural networks. She has published ten research papers in International conferences and National conferences. She has published several research papers in International journals and a book chapter.

Dr. P. Aruna was born in 1968 in India. She received her B.E. from Madras University, M. Tech from IIT Delhi and the Ph.D degree from Annamalai University. Presently she is working as a Professor in the Department of Computer Science and Engineering of Annamalai University. She has published 80 research papers in International Journals and Conferences and 26 research papers in National Journals and Conferences. She has twenty five years of teaching experience and sixteen years of research experience. She has published 3 book chapters. Her area of specialization includes Neural networks & Fuzzy systems, Data Mining and Image processing. She has guided 6 Ph.D scholars.

N. Puviarasan was born in 1963 in India. He received his M.S(Soft Sys) from BITS, Pilani, M.S (Engg.) from Annamalai University. Presently he is working as an Associate Professor in the Department of Computer Science and Engineering of Annamalai University. He has published 25 research papers in International journals and conferences and 13 research papers in National journals and conferences. He has twenty seven years of teaching experience and thirteen years of research experience. He has published one book chapter. His area of specialization includes Neural networks and Fuzzy systems, Data mining and Image Processing.