TM-SGTD: Text Mining Based Semantic Graph for Text Document Approach for Text Representation

Ashish Pacharne[1], Pramod S Nair[2], D Srinivasa Rao[3]
[1]Research Scholar, MIST, Indore, India
[2]Prof. of CSE Department, MITM, Indore,
[3]Assoc.Prof. of CSE Department, MITM, Indore, India
[1] ashishpne@gmail.com
[2] pramodsnair@yahoo.com
[3] sridas712@gmail.com

***Abstract*— Text representation is the essential step for the tasks of text mining. To represent the textual information more expressively, a kind of Text Mining based Semantic Graph approach is proposed, in which more semantic and ordering information among terms as well as the structural information of the text is incorporated. Such model can be constructed by extracting representative terms from texts and their mutually semantic relationships. The implementation of the proposed work is provided using the JAVA environment and python environment. Moreover, WordNet is showing relationship amongst word node. So that GEPHI tool is used to constructing more effectively semantic graph. Additionally the comparative performance is also compared with traditional. In order to compare the performance of the algorithms the memory consumption and time consumption is taken as stand parameters. The experimental results have proved the better performance of the proposed text information representation model in terms of its Time and Space complexity.**

Keyword- SVM, Semantic Graphs, POS Tagging, WordNet, Text representation, Text Mining, Graph Model, Semantic Networks

I. INTRODUCTION

Advances in digital technology and the World Wide Web have led to the increase of digital documents that are used for various purposes such as publishing and digital library. This phenomenon raises awareness for the requirement of effective techniques that can help during the search and retrieval of text. Nowadays, by using digital and computational techniques, we can store, manage and retrieve information automatically without any printed or hard copy of document. In addition of that in various applications automated text analysis or text mining played important role such as medical science, library management, social media and others. Typical tasks involved in these two areas include text classification, information extraction, document summarization, text pattern mining etc. [1]. Nowadays text is the most common form of storing the information. The representation of document is important step in the process of text mining. Hence, the challenging task is the appropriate representation of the textual information which will capable of representing the semantic information of the text [2]. In this work, we developed graph-based document model which is leverage valuable knowledge about relations between entities. Hence the work is intended to deliver a mechanism for constructing a semantic graph of text documents.

A. Semantic Graph

The data structure we will focus on is the semantic graphs. Semantic graphs are appropriate to represent the semantically information in their nodes, i.e., they carry semantic information on their nodes and edges. A semantic graph is a type of linkage of the different objects where nodes represent objects (e.g., persons, papers, organizations, etc.) and links (or edges) represent binary relationships between those objects (e.g., friend, citation, authorship, etc.). A semantic graph is a powerful representation structure which can encode semantic relationships between different types of objects. The edge relation information provides us the information of how the two different object nodes are connected to each other and their meaning. These graphs encode relationships as typed link between a pair of typed nodes. These semantically structured graphs are also called a relational data graph or an attributed relational graph. Indeed, semantic graphs are very similar to semantic networks and multi-relational networks (MRNs) used in artificial intelligence and knowledge representation [3].

B. Classification of Graph-based Model

The representation of a graph G is given using the tuple set $G = \{V, E\}$ where V is used for representing the nodes and E provides the edges of graph G. in the similar manner when the text is need to be presented using graph the vertices or node of graph represents the domain, subject or a word and the relationship among the nodes or these subjects or words are provided by the edges. We can classify the graphs in detail according to the node representation and the edge representation.

1.1 Node Representation Method

Nodes of graph G demonstrate the valuable components of the text such as words, subjects, domains, sentences, and paragraphs. All these components are used for representing the concepts or these can be considered as semantic components. By the definition of graph model, a component is represented by node, or can be indicate more than two components. If a node represents one component, it is called homogenous representation. If a node represents more than two components, it is called heterogeneous representation. Therefore in order to find the influence of the subject for other subject either graphs is directed or the graph used with labels. In this context the text representation graphs are used with some kinds of weights. Therefore the text graph representation supports both the techniques weighted or un-weighted according to requirements.

1.2 Homogenous representation vs. Heterogeneous Representation

Basically in homogenous representations a word is always represented as nodes [4]. In different places such as co-occurrence of a word are commonly expressed using graphical notations. Here the co-occurrence denotes that word appears more than once in a document or subject. Thus the words can be represented with some edge. Not only in this context the graphical notation of text also supports the grammatical associations among the words or semantic similarities in any homogeneous text graph representation. Since this representation is simple, the cost for building the model and analysis is low. The key advantage of this technique is that any existing graph representation algorithm can be used without modifications. In some researches, they also used homogenous representations, in which sentences, paragraphs or concepts are represented as nodes [5].

1.3 Weighted and Unweighted

In weighted representations, weighted value has been assigned in each node. On the other hand in un-weighted representations the nodes are not used with weights. Most researches assume weighted nodes which indicate the importance of the node in the graph. In order to evaluate the weight of nodes, some researches, for example PageRank, exploited the number of edges, the weights of edges, or the weight of nodes which are connected by the edge.

II. LITERATURE SURVEY

This section provides the recently made efforts and contributions to Design the techniques of semantic graph for text documents. Therefore, different articles and research papers are included in this section.

In this paper, ***Jinghua Wang et al. [6]*** introduced language network and described three kinds of networks. Keyword extraction is an important technology in many areas of document processing. In particularly, a keyword extraction algorithm based on language network and PageRank is proposed. Firstly a semantic network for a single document is build, then Pagerank is applied in the network to decide on the importance of a word, finally top-ranked words are selected as keywords of the document. The algorithm is tested on the corpus of CISTR, and the experiment result proves practical and effective.

In this paper ***Chuntao Jiang et al. [7]*** portrayed a chart based way to deal with record arrangement. The chart portrayal offers the preferred standpoint that it takes into consideration a substantially more expressive archive encoding than the more standard sack of words/expressions approach, and subsequently gives enhanced arrangement exactness. Report sets are known as a sets by which a weighted graph mining calculation is to remove visit sub-graphs, which are then additionally prepared to deliver highlight vectors (one for every record) for characterization. Weighted sub-graph mining is utilized to guarantee grouping adequacy and computational productivity; just the most noteworthy sub-graphs are removed. The approach is approved and assessed utilizing a few famous characterization calculations together with a certifiable printed informational index. The outcomes exhibit that the approach can outflank existing content grouping calculations on some dataset. At the point when the measure of dataset expanded, additionally handling on extricated visit elements is fundamental.

Programmed watchwords extraction is the errand to recognize a little arrangement of words, key expressions, catchphrases, or key fragments from a report that can portray the importance of the archive. Watchwords are helpful instruments as they give the most brief rundown of the archive. ***Vishal Gupta et al. [8]*** focuses on Automatic catchphrases extraction for Punjabi dialect content. It incorporates different stages like evacuating stop words, Identification of Punjabi things and thing stemming, Calculation of Term Frequency and Inverse Sentence Frequency (TF-ISF), Punjabi catchphrases as things with high TF-ISF score and title/feature highlight for Punjabi content. The extricated watchwords are particularly useful in programmed ordering, content outline, data recovery, arrangement, bunching, subject discovery and following and web looks and so forth.

In this paper, ***Marina Litvak et al. [9]*** present and think about between two novel methodologies, regulated and unsupervised, for recognizing the watchwords to be utilized as a part of extractive synopsis of content reports. Both this methodologies depend on the chart based syntactic portrayal of content and web records, which improves the conventional vector-space display by considering some auxiliary report highlights. In the managed approach, they prepare grouping calculations on a compressed accumulation of reports with the motivation behind instigating a watchword distinguishing proof model. In the unsupervised approach, they run the HITS calculation on archive charts under the suspicion that the top-positioned hubs ought to speak to the report watchwords. Our tests on a gathering of benchmark synopses demonstrate that given an arrangement of compressed preparing reports, the administered grouping gives the most elevated catchphrase distinguishing proof precision, while the most astounding F-measure is come to with a straightforward degree-based positioning. What's more, it is adequate to perform just the primary emphasis of HITS as opposed to running it to its joining.

Chien-Liang Liu et al. [10] proposed semi-administered grouping technique called Constrained PLSA to bunch labeled reports with a little measure of named archives and uses two informational collections for framework execution assessments. The primary informational collection is a record set whose limits among the groups are not clear; while the second one has clear limits among bunches. This review utilizes modified works of papers and the labels commented on by clients to bunch records. Four blends of labels and words are utilized for highlight portrayals. The trial comes about show that the greater part of the strategies can profit by labels. Nonetheless, unsupervised learning strategies neglect to work appropriately in the informational collection with uproarious data, however Constrained-PLSA works legitimately. In numerous genuine applications, foundation information is prepared, making it fitting to utilize foundation learning in the grouping procedure to make the adapting all the more quick and compelling.

III.PROPOSED WORK

This section includes the introduction of the proposed work and contribution made during the study. In addition of that the proposed model is also demonstrated with their working.

A. Problem Domain

Technology platforms are becoming increasingly more capable every day of interpreting and responding to domain specific problem of text that not practically to show leniently. Ontologies of the words can help represent the relationships between entities such that they can be used to improve the accuracy and effective representation of the system at meeting its users' information needs. Simultaneously, when user's information or any of generalized databases is highly lengthy comparable to generate the effective way to show word layout in graphical form is not possible to construct semantic graph. On the other hand, lingual words of the text word length are limited and it increases word length in databases. For that there is need to design a effective representation model of semantic based model of ontology based with co-occurrence of the word.

In most of the Ontologies are created by experts manually, therefore the design, development and maintenance or updating needs significant effort and time. To combat this, ontology learning systems, which attempt to automatically learn relationships from a domain and then map them into ontology, are becoming more prevalent

B. Methodology

For the refinement and advancement of traditional graph scenario here we present, a proposed Text Mining based Semantic Graph approach for text document, working flow that shows how the text data can be effectively depicts in a graphical view:

Description:

For constructing semantic graph for text documents is demonstrate in figure 1. In this, firstly, we take an input data in a textual form. Here we take a data from different news papers. After reading input file, secondly, pre-process the loaded file using normalization and stemming. Pre-processing is a process that identifies and removes the noisy content from the input data file for learning the procedure.

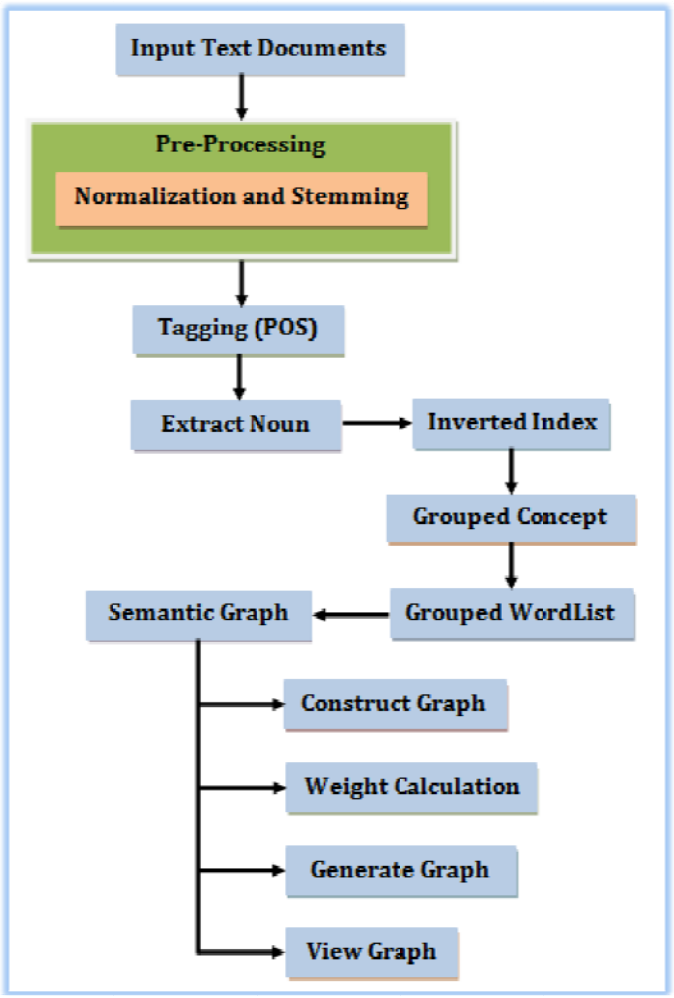

Fig.1. Proposed TM-SGTD Approaches

According to its name that process is applied before implementing the algorithm on the actual data. In the document mining the nature of pre-processing can be different from the other structured data mining techniques. In this phase we pre-process the data using normalization and stemming process. Normalization is a process that converts a list of words to a more uniform sequence. This is useful in preparing text for later processing. By transforming the words to a standard format, other operations are able to work with the data and will not have to deal with issues that might compromise the process. For example, converting all words to lowercase will simplify the searching process. For example, Nature, TEMPLE all will transform into "temple". Similarly, Stemming is a process in which the variant word forms are mapped to their base form. It is among the basic text pre-processing approaches used in Language Modeling, Natural Language Processing, and Information Retrieval applications. In this stemming process consider two phase, Firstly, we have consider WordNet Lemmatization. Lemmatization is the process which creates the set of lemmas of a lexical database. It is conceived as starting from text-words found in a corpus and leading to lemmas heading dictionary entries. For example, if we take a word "BOOKS", it will transform into "BOOK", but if we use it for "BOOKING" the output will be "BOOKING" because, the word "BOOKING" is independently defined in WordNet as verb and it has its self sense id. The second step of stemming process includes the application of the regular expression stemmer that is used in very specific cases. We are stemming just "ing" from word after this "BOKING" will be "BOOK".

After Pre-processing, we get refined text data on that we use POS tagger for tagging the data. The Part-Of-Speech Tagger is also termed as POS Tagger. That is a kind of software that accepts input as text and parse the text data into parts of speech information. This process is performed for entire words that are available in text input. The information from the text is recovered in terms of noun, verb, adjective, etc. POS Identifies the parts of speech to extract from the input text. Furthermore, we extract all the nouns from the tagged data. The tagged data is a combination of words and corresponding part of speech. On the extracted nouns, we apply indexing using inverted index. The main task of inverted index is to search keywords from documents. Here, inverted index match the keywords between two text data and retrieved most appropriately matched words. After

indexing, we prepare, grouped concept for the indexed words. All grouped concept make a grouped word list using Wordnet. The output of this step a wordlist which can be represented as $Wlist = \{w_1, w_2, \ldots . w_i\}$.

In the next phase, we are treating these words like nodes of the graph and finding the relation between words using WordNet. It is a significant to note that WordNet does not capture suitable noun like place or a person's name so these words will be treated like a single node and we can't find any association between these words.

Finally, we construct semantic graph using given algorithm. As well to constructed semantic graph, now calculate weight of each node of the graph. There is a different weight for each node relation that shows of each node is unique to each others. Significantly, put that nodes which have a highest node weight and generate a graph file corresponding analysis the semantic graph.

C. Proposed Algorithm

In the previous section the entire system design and proposed system architecture is demonstrated. In this given system model text based semantic graph modules are implemented for accurately constructing graph. Therefore in order to demonstrate the entire data process with their training and testing phase are used to demonstrate TM-SGTD. Table I shows the proposed algorithm of text representation:

TABLE I. Proposed Text Mining based Semantic Graph method for Text Document Algorithm

Input: Text Documents T_D ***Output:*** Semantic Graph S_G
Process: **1:** R = ReadInputTextData (T_D) **2:** Preprocessed = Normalization & $steamming$ (R) **3:** TaggedData = PartofSpeechTagging (Preprocesses Data) **4:** Nouns = getNouns(TaggedData) **5:** Indexing = InvertedIndex(Nouns) **6:** GroupedConcept = getConcept (Indexing) **7:** $Grouped$WordList = getList (GroupedConcept) **8:** SemanticGraph = ConstructGraph (WordList) **9:** Return S_G

Graphs are a well-studied class of data structures used to model relationships (edges) between entities (nodes). Semantic Graph in general and ontologies specifically, model a domain by defining how different entities within the domain are related. In order to make understandable the text data can be visualized using the graphs, in this representation the text data is modeled in terms of nodes and edges. The edges represent the relationships between nodes. This process either performed manually or by some domain expert or automatically done by ontology based system. Because building such a semantic graph base typically requires explicitly modeling nodes and edges this ahead of time.

A graph representation of text can have any combination of nodes and edges towards other nodes. That materializes to demonstrate relationships between nodes. The knowledge representation using graphs demonstrate a number of advantages such as their development and automatically creation using real-world data, demonstration of different combination of nodes and edges include the pre-existing nodes or domain, and a complete modeling of the semantic relationships between all entities in a domain and dynamically traversal of the graph.

IV. RESULT ANALYSIS

A. Time Consumption

Figures must be numbered using Arabic numerals. Figure captions must be in 8 pt Regular font. Captions of a single line (e.g. Fig. 2) must be centered whereas multi-line captions must be justified (e.g. Fig. 1). Captions with figure numbers must be placed after their associated figures, as shown in Fig.1.

$$\text{Time Consumed} = \text{End Time} - \text{Start Time}$$

The time consumption of the proposed algorithm is given using figure 2 and table II.

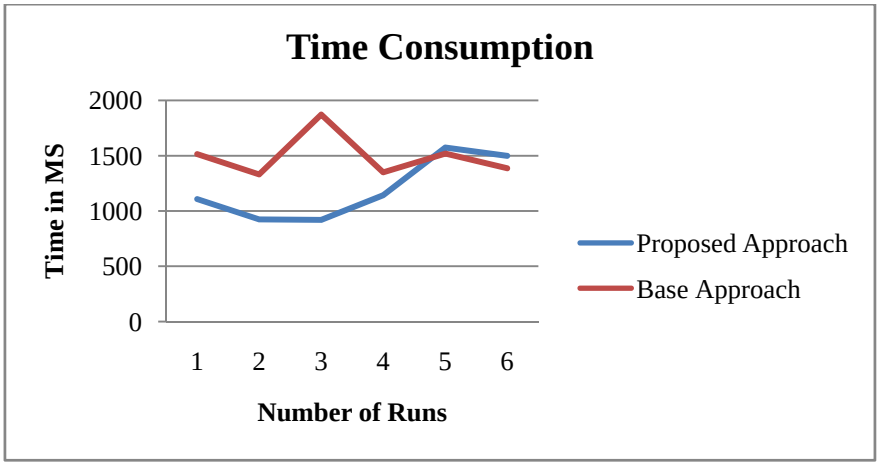

Fig.2. Time Consumption

In order to show the performance of implemented approaches the time consumption is reported in figure 2 and table II. In this diagram the X axis shows the different experiments on which different values generated and the Y axis shows the amount of time consumed for processing the algorithm with respective of input data file. Additionally the performance of proposed TM-SGTD is given using blue line and traditional approach depicts using orange line. According to the given results the proposed system consumes less time as compared to other traditional algorithm. Here the time required to matching the text keywords using indexing. Additionally the results shows the amount of time consumed is depends on the amount of data provided for execution of algorithm. But the respective performance of the system shows their effectiveness over the traditional algorithm. Moreover, while implementing TM-SGTD is representing text data over the graphical form efficiently.

TABLE II. Time Consumption

Number of Runs	Proposed Approach (Time in Millisecond)	Base Approach (Time in Millisecond)
1	1108	1515
2	925	1331
3	920	1871
4	1142	1350
5	1573	1518
6	1498	1386

B. Memory Consumption

The memory consumption shows the amount of main memory required to process the algorithm with input amount of data to be processed. That is also known as the space complexity of algorithm. To compute the memory consumption, the following formula is used.

$$\text{Memory Consumed} = \text{Total Memory} - \text{Free Memory}$$

TABLE III. Memory Consumption

Number of Runs	Proposed Approach (Memory in KB)	Base Approach (Memory in KB)
1	185577	269275
2	178878	241116
3	167073	140955
4	289200	312628
5	156899	184519
6	145235	195423

The figure 3 and table III shows the memory consumption or space complexity of the system with increasing the number of runs for text documents. The unit of experiments performed with the text dataset is given using X axis and Y axis describes memory requirements of the algorithms during experimentation. The amount of memory requirement is measured in terms of kilobytes (KB). According to the experimented results the amount of memory is not similar as much higher and not more fluctuating. So, this graph shows the proposed system is not highly consumed memory other than traditional system. When the semantic graph constructed system take more space for evaluating nodes and edges but traditional approach have large memory overhead to for processing data.

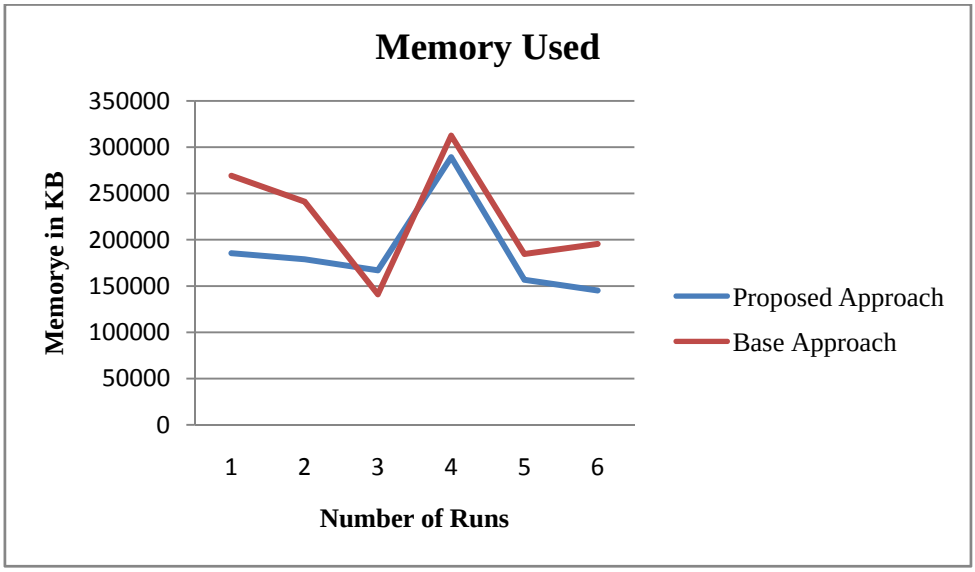

Fig.3. Time Consumption

IV. CONCLUSION

The rapid developments of modern techniques have enabled a large amount of text data to be published on the web. However, effective representation of text for various text mining tasks is still an open problem. In popular and classical text representation models of text analysis and visualization the graphs are developed using domains such as single words or phrases, additionally that is assumed that all terms are treated as independent units of graphs.

In this research work, we present a new idea for mining documents by exploiting semantic information of their texts. We have proposed, *TM-SGTD* i.e. *Text Mining based Semantic Graph approach for Text Document*, a novel approach which present the semantic graph that constructed using text data. The approach is a formal semantic representation of linguistic inputs is introduced and utilized to build a semantic representation scheme for documents. An approach to text representation using a semantic graph has been described. The graph representation of text allows both the structure and content of documents to be represented.

V. FUTURE WORK

The primary goal of the proposed work is achieved successfully. In further for the more improvements and their different application feasibility the following suggestions are made for extensions.

- In near future work use the semantic similarity based technique and word sense disambiguation to improve connectivity and relations between nodes.
- In near future work use the semantic similarity based technique and word sense disambiguation to improve connectivity and relations between nodes.
- Another future direction is to investigate the usage of WordNet to extract the synonyms, hypernyms, and hyponyms and their effect on document clustering, categorization, and retrieval results, compared to that of traditional methods.

REFERENCES

[1] Andreas Hotho, Andreas Nrnberger, and Gerhard Paa, A brief survey of text mining, LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 2005.
[2] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo, SEM 2013 shared task: Semantic textual similarity. In Proc. of *SEM-2013, pages 32–43, 2013.
[3] Barthélemy, Marc, Edmond Chow, and Tina Eliassi-Rad. "Knowledge Representation Issues in Semantic Graphs for Relationship Detection." AAAI Spring Symposium: AI Technologies for Homeland Security. 2005
[4] Yadav, Chandra Shekhar, Aditi Sharan, and Manju Lata Joshi, "Semantic graph based approach for text mining", Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on, IEEE, 2014.

[5] Chang, Jae-Yong, and Il-Min Kim, "Analysis and evaluation of current graph-based text mining researches", advanced Science and Technology Letters 42 (2013): 100-103.
[6] Liu, Jianyi, and Jinghua Wang, "Keyword extraction using language network", International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE 2007, IEEE, 2007.
[7] Jiang, Chuntao, et al. "Text classification using graph mining-based feature extraction." Knowledge-Based Systems 23.4 (2010): 302-308.
[8] Gupta, Vishal, and Gurpreet Singh Lehal, "Automatic keywords extraction for Punjabi language", International Journal of Computer Science Issues 8.5 (2011).
[9] Litvak, Marina, and Mark Last, "Graph-based keyword extraction for single-document summarization", Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, Association for Computational Linguistics, 2008.
[10] Liu, Chien-Liang, et al "Clustering tagged documents with labeled and unlabeled documents."Information Processing & Management 49.3 (2013):596-60.

AUTHOR PROFILE

Ashish Pacharne received BE degree in electronics and communication engineering and is currently working towards the ME degree in the Department of Information Technology at Medicaps Institute of Science and Technology, Indore Madhya Pradesh. His Research interest includes data mining, text mining, social media mining.

Pramod S Nair received B.Tech, M.Tech and Ph.D degree in Computer Science. His research interests include Data Mining, Web Mining Internet applications, Information Retrieval and Knowledge Discovery. He has published many research papers in international conferences and journals. He is also associated with review panels of different conferences and journal of international repute.

D Srinivasa Rao M.Tech, (Ph.D) is working as an Associate Professor in the Department of Computer Science & Engineering at Medicaps Institute of Technology and Management, Indore, Madhya Pradesh, India. He has 20 years of teaching experience. His area of interest in Adhoc Networks, Distributed Systems, Network Security, Image Processing & Data Mining. He has guided more than 60 Post Graduate Students. He has published 2 books and 25 papers in international journals. He presented 2 papers in National Conferences, 1 paper in International Conference and has attended 35 National Workshops / FDP / Seminars etc. He is a life member of Professional Society like ISTE.