ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

An Automatic Topic Summarization Using Content Anatomy

[I]R.Sinduja, [II]Dr. S.Senthamarai Kannan
[I]PG student, [II]Professor
[I,II]Dept. of CSE, Sethu Institute of Technology, Affiliated to Anna University, Kariapatti, Tamilnadu, India

Abstract
A topic is defined as a seminal event or activity along with all directly related events and activities. It is represented by a chronological sequence of documents published by different authors on the Internet. We define a task called topic anatomy, which summarizes and associates the core parts of a topic temporally so that readers can understand the content easily. The proposed topic anatomy model, called TSCAN, derives the major themes of a topic from the eigenvectors of a temporal block association matrix. Finally, the extracted events are associated through their temporal closeness and context similarity to form the evolution graph of the topic While current technologies are efficient in searching for appropriate documents to satisfy keyword search requests, users still have difficulty assimilating needed knowledge from the overwhelming number of documents. The situation is even more confusing if the desired knowledge is related to a temporal incident about which many independent authors have published documents based on various perspectives that, considered together, detail the development of the incident.

Keywords
Text mining, event segmentation, theme generation, topic summarization, content anatomy

I. Introduction

A text summarizer strives to produce a condensed representation of its input, intended for human consumption. It may condense individual documents or groups of documents. Text compression, a related area, also condenses documents, but summarization differs in that its output is intended to be human-readable. The output of text compression algorithms is certainly not human-readable, but neither is it actionable the only operation it supports is decompression, that is, automatic reconstruction of the original text. As a field, summarization differs from many other forms of text mining in that there are people, namely professional abstractors, who are skilled in the art of producing summaries and carry out the task as part of their professional life

Text mining is the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. In the beginning, Information Retrieval (IR) provided many term-based methods to solve this challenge. Just as data mining can be loosely described as looking for patterns in data.

To promote research on detecting and tracking incidents from Internet documents, the Defense Advanced Research Projects Agency (DARPA) initiated the Topic Detection and Tracking (TDT) project. The project defines a topic as "a seminal event or activity, along with all directly related events and activities." Its goal is to detect topics automatically and track related documents from several document streams, such as online news feeds.

II. Related Works

Earlier works on summarization methods has been expansively studied in text mining communities for many years. A Variety of efficient algorithm are used. Such as, forward method, backward method, SVD method, K-means method, Temporal summary (TS) method, frequent content word method (FCW), TSCAN has been proposed. The main problem in text mining is finding the closed pattern.

A. SVD method

SVD [6] method uses a particularly efficient algorithm for singular value decomposition that can handle even very large input matrices (of word counts and documents). Assume matrix A represents an m x n word occurrence matrix where m is the number of input documents (files) and n the number of words selected for analysis. SVD computes the m x r orthogonal matrix U, n x r orthogonal matrix V, and r x r matrix D, so that A = UDV', and so that r is the number of eigenvalues of A'A.

For most Text Mining problems, the SVD will be entirely appropriate to use. Without a data reduction technique, there will be more variables (terms) available than one can use in a data mining model. Some method must be applied to select an appropriate set from which a text mining solution can be built. Unlike term elimination, the SVD technique allows one to derive significantly fewer variables from the original variables. There are some drawbacks to using the SVD, however. Computationally, the SVD is fairly resource intensive and requires a large amount of RAM. The user must have access to these resources in order for the decomposition to be obtained SVD method is used to composes the summaries by extracting the blocks with the largest entry value in singular vectors. SVD method is using graph based summarization method. Note that the result derived by the SVD method is identical to that of the graph based summarization method.

B. K-means method

The k-means algorithm [19] is used for efficiency in clustering large data sets. However, working only on numeric values prohibits it from being used to cluster real world data containing categorical values. The k-means algorithm (MacQueen, 1967; Anderberg, 1973), one of the mostly used clustering algorithms, is classified as a partition or non-hierarchical clustering method. It can be used to cluster texts. K-means algorithm is an algorithm to partition and classify the data based on attributes or features in to k-number of groups.

The k-means algorithm has the following important properties:

1. It is efficient in processing large data sets.
2. It often terminates at a local optimum (MacQueen, 1967; Selim and Ismail,1984).
3. It works only on numeric values.

The K-means method which compiles summaries by selecting the most salient blocks of the resulting K clusters. This method's

performance depends on the quality of the initial clusters. In this experiment, to ensure fair comparison of the K- means method, which provide the best result from 50 randomly selected initial clusters for evaluation.

C. Temporal summary (TS) method

Temporal summary method[8] is one of the summarization methods for content discovery. The temporal summary (TS) method take on the useful2 and novel1 techniques proposed by the authors to compute the in formativeness score of a topic block. we do not take on the novel2 technique because the authors have shown that the performance difference between using novel1 and using novel2 is not significant. In addition, novel2 requires a training corpus to derive an appropriate number of clusters (i.e., parameter m), but the training corpus is not available.

D. Frequent content word method (FCW)

Frequent content method[9] is used to construct the summaries by using selecting the block with frequent terms This method's performance is comparable to that of state-of-the- art summarization methods. In addition, we adopt Nenkova et al.'s context adjustment technique to increase the summary diversity.

IV. Proposed Approach

TSCAN method stands for Topic Summarization and Content Anatomy (TSCAN), which organizes and summarizes the content of a temporal topic by using set of documents. TSCAN models the documents as a symmetric block association matrix, in which each block is a portion of a document, and treats each eigenvector of the matrix as a theme embedded in the topic. The eigenvectors are then examined to extract events and their summaries from each theme.

A. Crawling

We use the WebCrawler for retrieve the document. In a web page we retrieve the contents by the way of web crawler. We get the webpage URL and using to web crawler and it retrieve the documents. From this crawling, we are getting topic and the inbound links for the particular topic.

B. Extraction

Extracting blocks from inbound links in the web sites. In this block extraction extracting blocks and image for the topic. We obtain more number of blocks from the web sites.

After extracting blocks, Events are extracted from the blocks. Set of events are extracted for every blocks.

C. Matrix Calculation

Calculating Eigen Values and Eigen Vector for every event from the blocks. By calculating the Eigen vector for events, we are getting a effective events for the particular topic from the peak values of the eigen vector. Eigen vector with peak value is taken as a worthy event

The matrix $A=B^TB$, called a block association matrix, is an nxn symmetric matrix in which the (i, j)-entry (denoted as ai,j) is the inner product of columns i and j of the matrix B. As a column of B is the term vector of a block, A represents the inter-block association. Entries with a large value imply a high correlation between the corresponding pair of blocks. A theme of a topic is regarded as an aggregated semantic profile of a collection of blocks.

Theme can be represented as a vector vof dimension n, where each entry denotes the degree of correlation of a block to the theme. To acquire appropriate themes of the topic, the theorem of symmetric matrices is employed. The matrix A can be represented as follows

$$A=VDV^{-1}\text{-}VDV^T$$
$$=[v_1,.....,v_n][d_{1,1}e_1,.....d_{r,r}ed_r,0e_{r+1}.....e_n]V^T$$
$$=[d_{1,1}v_1,.....,d_{r,r}v_r,0v_{r+1},.....0v_n][v_1,.....v_n]^T$$
$$= d_{1,1}v_1v^T+...+d_{r,r}v_r\,v^T+0v_{r+1}0v_{r+1}^T+0v_nv_n^T$$

A and di,i is its corresponding eigenvalue. i.e. the symmetric matrix A can be decomposed into the sum of n matrices spanned by its eigenvectors. We take the first L (L<r) significant eigenvectors of A as the themes of the topic.

The inter-block association approximated by the selected themes can be represented as follows:

$$A \approx d_{1,1}v_1v^T + \ldots\ldots + d_{L,L}v_Lv^T$$
$$=[v_1,.....v_L][d_{1,1}e_1,....,d_{L,L}e_L][v_1,.....v_L]^T$$
$$=v_LD_Lv$$

Eigenvalues of A. i.e. the interblock association of topic can be approximated by selecting a certain number of themes with significant eigenvalues. As the eigenvectors of A are orthogonal to each other, the produced themes tend to be unique and descriptive.

D. Event Segmentation

By using Eigen vector, we merge close events and we prune small events for every locks. In this event segmentation, every close event is merged together for every block in a topic.In the R-S algorithm, every block in an eigenvector has an energy value, which is defined as follows:

$$eng(i,j)=\frac{1}{H}\sum_{h=-(H-1)/2}^{(H-1)/2}[v_{i+h,j}]^2,$$

where eng (i,j) is the energy of a block I in a theme j, and H specifies the length of a sliding window used to smooth and aggregate the energy of a block with that of its neighbourhood.

E. Summarization

Event Summarization, which merges every closely matching blocks thus forming a summarization of the topic.

Finally, Topic story is detected and aligned in chronological order.

V. Performance Evaluation

Traditionally, performance evaluations in information retrieval depend on annotated benchmarks. But there are no official benchmarks and metrics for the study of topic anatomy. So we compare the performance of several summarization methods, as it is a common practice in summarization evaluation, and evaluate the topic evolution graphs generated by TSCAN to demonstrate the model's capability.

The experiments employed the official TDT corpus where 26 news topics, each containing more than 20 documents, were selected for performance evaluations. Each topic document is partitioned into blocks of sentences. Each block has 3 sentences to ensure that it contains a complete sentence. The parameter Liscritical to the quality of detected themes. Larger the number of themes selected, the better the approximation will be. For summarization

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

comparisons, the evaluations are performed with L=1 to 10 to show the influence of themes on summarization performance. Also the parameter H and the temporal similarity threshold are set to 7 and 0.3, respectively.

Table 1: Statistics of Evaluated Topics

Number of topics	26
Number of news documents	1211
Average number of documents per topic	46.6
Number of sentences	32739
Average number of sentences per topic	1259.2

VI. Summarization Evaluations

We compare the summarization performance of TSCAN with the following four summarization methods:

1. 1.The forward method, which generates summaries by extracting the initial blocks of a topic.
2. 2.The backward method, which extracts summaries from the end blocks of a topic.
3. 3.The SVD method, which composes summaries by extracting the blocks with the largest entry value in singular vectors.
4. 4.The K-means method, which composes summaries by selecting most salient blocks of the resulting K clusters.

The summarization evaluation procedure is as follows: For each L, we first apply TSCAN to each topic to extract a set of blocks as the topic summary. We use the compared methods' algorithms, and then produce summaries of the same size (in terms of the number of blocks) as those generated by TSCAN. The compression ratios for summaries of L produced by the compared methods In sum, the compression ratios of the evaluated summaries are high and at least 90% of the topic's contents are omitted.

Table 2. The Number of Blocks Accumulated for Summaries

	$w=1$			$w=3$			$w=5$		
L	$H=5$	$H=7$	$H=9$	$H=5$	$H=7$	$H=9$	$H=5$	$H=7$	$H=9$
1	7.3 (99%)	12.7 (98%)	11.7 (98%)	7.5 (98%)	8.7 (97%)	7.6 (98%)	5.1 (98%)	5.5 (98%)	4.8 (98%)
2	11.0 (98%)	21.6 (97%)	19.2 (98%)	9.3 (97%)	12.5 (96%)	11.1 (97%)	7.1 (97%)	8.3 (96%)	7.3 (97%)
3	12.9 (98%)	28.0 (97%)	24.4 (97%)	10.7 (97%)	16.5 (95%)	14.3 (96%)	8.5 (96%)	11.0 (95%)	9.8 (95%)
4	14.7 (98%)	34.5 (96%)	30.0 (96%)	12.8 (96%)	20.7 (94%)	17.9 (95%)	10.2 (95%)	14.0 (94%)	12.4 (94%)
5	16.7 (98%)	41.6 (95%)	35.8 (96%)	13.9 (96%)	23.8 (93%)	20.6 (94%)	11.1 (95%)	15.8 (93%)	14.2 (93%)
6	17.7 (98%)	46.3 (95%)	39.8 (95%)	15.4 (95%)	27.8 (92%)	23.5 (93%)	12.3 (94%)	18.5 (91%)	16.5 (92%)
7	19.8 (98%)	52.7 (94%)	44.7 (95%)	16.4 (95%)	30.5 (91%)	25.8 (93%)	13.7 (94%)	21.2 (90%)	19.1 (91%)
8	21.0 (97%)	57.7 (94%)	48.9 (95%)	17.9 (95%)	33.8 (90%)	28.5 (92%)	14.9 (93%)	23.4 (89%)	21.2 (90%)
9	22.0 (97%)	62.3 (93%)	52.8 (94%)	19.1 (94%)	37.3 (89%)	31.5 (91%)	16.0 (93%)	25.7 (88%)	23.1 (89%)
10	22.9 (97%)	66.2 (93%)	56.0 (94%)	20.8 (94%)	41.2 (88%)	34.8 (90%)	16.7 (92%)	27.9 (87%)	24.8 (88%)

A. Summary-To-Document Content Similarity

Summary-to-document content similarity is defined as the average cosine similarity between an evaluated summary and topic documents, both of which are represented by TF-IDF term vectors. A high similarity score implies that the summary is representative of the topic and can effectively replace the original topic documents for various information retrieval tasks. Table 2 shows the micro average summary-to- document content similarity derived by the compared methods.

The comparison result shows, the TSCAN has a lower performance than the K-means summarization method on big summaries.

B. Rouge Evaluation

ROUGE is a recall-oriented summary evaluation metric used mostly in DUC contests (Document Understanding Conferences). It measures summarization performance by calculating the number of overlapping n-grams between an evaluated summary and a set of reference summaries. ROUGE scores 1 when the evaluated summary is consistent with the reference summaries; and 0 when the evaluated summary is off topic.

Therefore, we use ROUGE-1 (unigram overlapping) and ROUGE-2 (bi-gram overlapping) to evaluate the consistency of manual summaries derived by the compared methods. Tables 2 and 3 show the micro average performances of ROUGE-1 and ROUGE-2, respectively.

C. Scalablity and Time Comparision

We evaluated the execution time of the compared summarization methods on an AMD AthlonTM 64 Processor 3200++ PC with the Windows XP Service Pack 3 operating system and a 2 GB main memory.

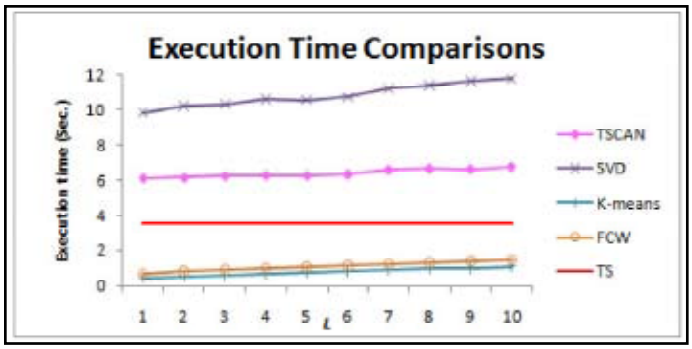

Fig. 1. Comparision of execution time

For each method, we recorded the time required to generate the summaries of the 26 evaluated topics under a specific parameter setting. However, due to space limitations, we only show the average execution times of the methods under all parameter setting We do not show the execution times of the forward and backward methods because they do not need to weight each topic block to compile topic summaries.

Therefore, their respective execution times are constant and irrelevant to the parameter settings. FCW is an iterative summarization method. In each iteration, it computes the weight of every block based on a topic model; therefore, its time complexity is O(n). K-means is also a linear algorithm[3].

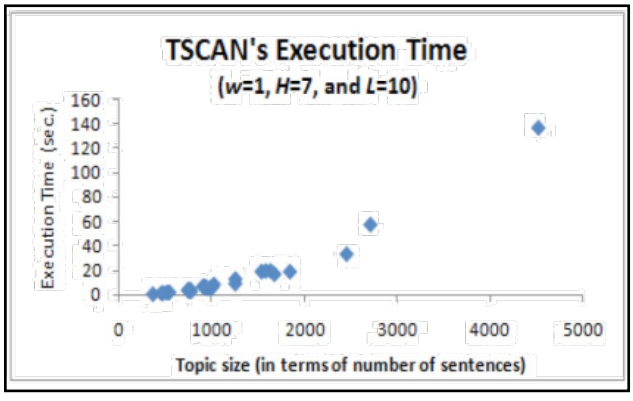

Fig. 2. The scalability of TSCAN

Although the TS method computes the weight of topic blocks linearly, it needs to examine the content of previous blocks to compute a block's novelty; therefore, its time complexity is O(n2). The SVD method and our method spend much of the computation time calculating eigenvectors. We employ MATLAB to calculate a matrix's eigenvectors. The time complexity is O(n2I2), where I is the number of eigenvectors to be computed [18]. linear summarization methods (i.e., Kmeans and FCW) run faster than the other methods.

In terms of time complexity, the eigen-based methods (i.e., the SVD method and our method) run slower than the TS method. However, we observe that when w ¼ 3 and w ¼ 5 our method runs faster than the TS method. We believe that the longer execution time of the TS method is due to program implementation issues, as MATLAB is a commercial software package implemented by experienced programmers. Except for the TS method, the execution time of the compared methods generally increases as the size of the summary (i.e., L) increases. For the SVD method and TSCAN, a large L means that the methods need to examine a lot of eigenvectors to compile topic summaries; therefore, the execution time increases. For the K-means method and the FCW method, a large L increases the number of clusters (i.e., K) and the number of iterations required to extract summary blocks with frequent terms, respectively. Hence, the methods' execution times also increase. It is noteworthy that the TS method's execution time is irrelevant to the size of the summary. This is because the method must weight all topic blocks irrespective of how many summary blocks are required to compile a topic summary.

To evaluate the scalability of TSCAN, we show the time it requires to generate the summaries of the 26 topics in.For the largest topic which consists of 128 topic documents and 4,527 topic sentences, TSCAN takes approximately 2 and half minutes to construct the topic summary. In practice, the computation time might be less than the time required to crawl the topic documents. Thus, the proposed method is feasible

Table 3 Microaverage performance of ROUGE-2

	TSCAN	Forward		Backward		SVD		K-means	
L	R-2	R-2	IMP	R-2	IMP	R-2	IMP	R-2	IMP
1	0.139	0.088	57.7%	0.062	125%	0.096	44.6%	0.098	41.7%
2	0.144	0.107	34.8%	0.082	76.7%	0.117	23.2%	0.121	18.7%
3	0.154	0.115	33.9%	0.093	65.1%	0.124	24.6%	0.134	14.9%
4	0.160	0.127	25.7%	0.112	43.1%	0.129	23.9%	0.142	12.1%
5	0.161	0.133	21.1%	0.119	35.9%	0.135	19.8%	0.143	12.7%
6	0.169	0.147	14.8%	0.129	31%	0.150	12.6%	0.153	10.1%
7	0.170	0.150	13.3%	0.134	27%	0.151	13.1%	0.158	8.1%
8	0.174	0.155	12.3%	0.144	20.8%	0.156	11.8%	0.165	5.5%
9	0.177	0.164	7.7%	0.153	15.4%	0.162	8.8%	0.169	4.8%
10	0.180	0.173	4.2%	0.167	8.2%	0.172	5.1%	0.179	1%

Legend: R-2: the micro average ROUGE-2 performance

VII. Performance Review

To summarize, we evaluate the summarization performance of TSCAN in terms of content coverage, content coherence, consistency with expert-composed summaries, and execution time. The experiment results show that, as well as covering the core parts of evaluated topics, our summaries are content-coherent and consistent with expert- composed reference summaries. The quality of our summaries is better than that of many well-known summarization methods, especially when the compression ratio is high.

For example, when L¼1, TSCAN outperforms the compared methods by 2.1 to 93.1 percent for SDCS, by 26.8 to 361.4 percent for APSBS, by 7.2 to 85.9 percent for ROUGE-1, and by 7.2 to 163.1 percent for ROUGE-2 with a 95 percent confidence level based on a one-tailed paired t- test. These results demonstrate that TSCAN can select representative sentences earlier than the compared methods when compiling topic summaries. In resource-limited environments, such as when the network bandwidth is low, this property helps users capture key information about a topic. While our method's execution time is longer than that of many well-known summarization methods, the time required to compile the summaries of large topics is a few minutes at most; thus it is feasible for practical text few minutes at most; thus, it is summarization systems. The improvements over the com- pared methods also emphasize the importance of temporal information (i.e., events) in temporal topic summarization. In addition to providing summary diversity, topic summarization methods should summarize the developments of significant themes that produce content-coherent summaries and are consistent with human-composed summaries.

VIII. Conclusion and Future Enhancement

Publishing activities on the Internet are now so prevalent that when a fresh news topic occurs, autonomous users may publish their opinions during the topic's life span. To help Internet users grasp the gist of a topic covered by a large number of topic documents, text summarization methods have been proposed to highlight the core information in the documents. Most summarization methods try to increase the diversity of summaries to cover all the important information in the original documents. However, when the documents to be summarized are related to an evolving topic, summarization methods should also consider the temporal properties of the topic in order to describe the development of storylines.

In this paper, we have presented a topic anatomy system called TSCAN, which extracts themes, events, and event summaries from topic documents. Moreover, the summarized events are associated by their semantic and temporal relationships, and presented graphically to form an evolution graph of the topic. Experiments based on official TDT4topics demonstrate that TSCAN can produce highly representative summaries that correspond well to the reference summaries composed by experts.

In future instead of online news document, other unconstrained texts, such as blogs can be summarized.

References

[1] *J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic Detection and Tracking Pilot Study: Final Report," Proc. US Defense Advanced Research Projects Agency (DARPA) Broadcast News Transcription and Understanding Workshop, pp. 194-218, 1998.*

[2] *V. Hatzivassiloglou, L. Gravano, and A. Maganti, "An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering," Proc. 23rd Ann. Int'l ACM SIGIR Conf.Research and Development in Information Retrieval, pp. 224-231, 2000.*

[3] *C.D. Manning, P. Raghavan, and H. Schutze, "Introduction to Information Retrieval". Cambridge Univ. Press, 2008.*

[4] *Y. Yang, T. Pierce, and J. Carbonell, "A Study on Retrospective and Online Event Detection," Proc. 21st Ann.Int'l ACM*

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

SIGIR Conf. Research and Development in Information Retrieval, pp. 28-36,1998.

[5] D.M. Blei and P.J. Moreno, "Topic Segmentation with an Aspect Hidden Markov Model," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,pp. 343-348, 2001.

[6] Y. Gong and X. Liu, "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 19-25, 2001.

[7] T. Nomoto and Y. Matsumoto, "A New Approach to Unsupervised Text Summarization," Proc. 24th Ann. Int'l ACM SIGIR Conf.Research and Development in Information Retrieval, pp. 26-34, 2001.

[8] G. Erkan and D.R. Radev, "LexRank: Graph-Based Centrality as Salience in Text Summarization," J. Artificial Intelligence Research,vol. 22, pp. 457-479, 2004.

[9] A. Nenkova, L. Vanderwende, and K. Mckeown, "A Compositional Context Sensitive Multi-Document Summarizer: Exploring the Factors that Influence Summarization," Proc. 29th Ann. Int'lACM SIGIR Conf. Research and Development in Information Retrieval,pp. 573-580, 2006.

[10] J. Kleinberg, "Bursty and Hierarchical Structure in Streams," Proc.Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and DataMining, pp. 91-101, 2002.

[11] R. Nallapati, A. Feng, F. Peng, and J. Allan, "Event Threading within News Topics," Proc. 13th ACM Int'l Conf. Information and Knowledge Management, pp. 446-453, 2004.

[12] C.C. Yang and X. Shi, "Discovering Event Evolution Graphs from Newswires," Proc. 15th Int'l Conf. World Wide Web, pp. 945-946, 2006.

[13] A. Feng and J. Allan, "Finding and Linking Incidents in News,"Proc. 16th ACM Conf. Information and Knowledge Management,pp. 821-830, 2007.

[14] R. Swan and J. Allan, "Automatic Generation of Overview Timelines," Proc. 23rd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 49-56, 2000.

[15] T. Hofmann, "Probabilistic Latent Semantic Indexing," Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 50-57, 1999.

[16] C. Nicholas and R. Dahlberg, "Spotting Topics with the Singular Value Decomposition," Proc. Fourth Int'l Workshop Principles of Digital Document Processing, pp. 82-91, 1998.

[17] L.R. Rabiner and M.R. Sambur, "An Algorithm for Determining the Endpoints for Isolated Utterances," the Bell System Technical J.,vol. 54, no. 2, pp 297-315, Feb. 1975.

[18] J. Ruan and W. Zhang, "An Efficient Spectral Algorithm for Network Community Discovery and Its Applications to Biological and Social Networks,"Proc. Seventh IEEE Int'l Conf. Data Mining,pp 643-648, 2007

[19] T. Nomoto and Y. Matsumoto, "A New Approach to UnsupervisedText Summarization," Proc. 24th Ann. Int'l ACM SIGIR Conf.Research and Development in Information Retrieval, pp. 26-34, 2001

Sinduja.R received her B.Tech degree in Information Technology from Sree Sowdambika College Engineering,India in 2011. She currently pursuing her M.E degree in computer science and engineering from Sethu Institute of Technology. Her interest includes datamining,date warehousing.