The Content Summarization system

A.Pearline Divya[1], S.Leela[2], Asst. Prof
Department of Computer Science and Engineering, Karunya University
Coimbatore, India

Abstract- **The documents that are published in the internet are represented in the chronological sequence. Any interesting part of this document is called as topic. Topic is the sentence which captures the attention of the readers, and the core parts of the topic will be associated temporarily so that they would help the internet news readers to grasp its content easily. The content summarization model makes use of Eigen vectors and temporal block association matrix to generate themes of the topic. Further based on the construction of Eigen vectors, significant actions and summaries are generated from the themes. Finally Graph building is done based on the temporal closeness and contextual similarities between the actions. This would finally result in summarized content to the internet news readers. The summaries generated through our system proved to be superior in terms of information coverage and consistency when compared with other traditional methods.**

Keywords- text mining, language summarization, text analysis

I.INTRODUCTION

There are many other means to get information rather than traditional media. Internet documents play the vital role here. Everything became computerized, and even the day to- day's news. News that is provided in the internet are more in number and it is highly impossible for the readers to read through all the related documents. To improve research on detecting incidents and tracking related works, the Defence Advanced Research Projects Agency (DARPA) originated Topic Detection and Tracking (TDT) projects [6]. It provides techniques for detecting and tracking actions from several document streams. But TDT suffered many practical implication problems like segmenting stream of data into distinct stories, identifying those news stories to be the new event that is occurred for the first time and finding the Relationship between storylines provided a small number of News documents. Effective TDT helps in finding the topic and tracking all the related action but users cannot retrieve the exact details unless they read through many of the tracked documents. Hence there is an urgent need for summarization method to find the significant part of the detected topics. Finally to provide a pictorial representation that connects the significant part s of those detected topics. On considering these two facts, our proposed system- content summarization system, can summarize essential information about the topic in a chronological order.

Over the life span of the topic, the topic's content may change from one theme to another which signifies topic's development. Our approach deals with three major task which has to be followed in an orderly way. They are: Theme extraction, action segmentation and summarization and Graph building. By analyzing the intention variation of themes over a period of time, action segmentation and summarization process extracts topic actions and there summaries. Following Fig.1 shows a content summarization for the TDT4 [9] topic "president Dr. Abdul Kalam insist on mandatory education". The topic documents, experts found that there are three themes and four actions. The three themes that are generated are "Dr. Abdul Kalam's attitude towards education", "mandatory Education foes' opinion" and "commentators review on Dr.Abdul Kalam's decisions". The temporal properties are of important concern here. The content anatomy system models shows the internet documents as a block association matrix and considers each eigenvector of the matrix to be the theme embedded in the topic. From the themes generated, the actions would be filtered out. The experimental outcome of the TDT4 corpus shows that our system's summaries are highly representative and more consistent with high coverage of data.

II.RELATED WORKS

A. Text Segmentation

The text segmentation is the process of portioning the input text into non-overlapping blocks of data [15]. Many different techniques have been used to find the segregation between different sentences E.g. Naïve approach [25]. Similarly document subtopic identification is where a document will be provided as input and there by a subtopic of interest will be identified and paragraphs where as with topic segmentation; the identified actions of the themes have temporal property.

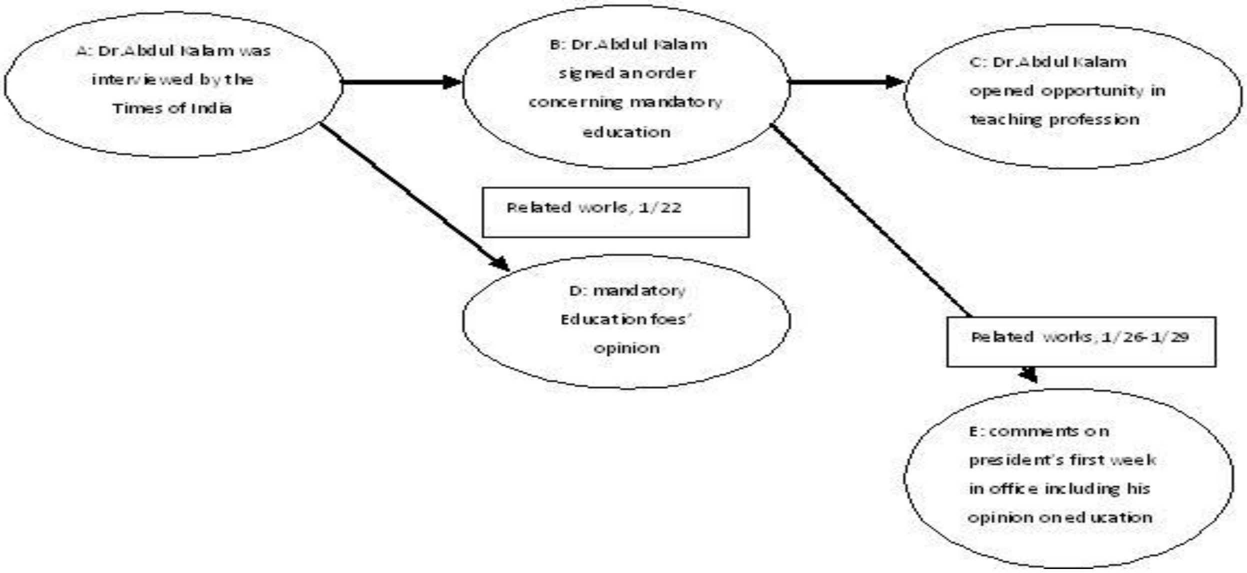
A: Dr.Abdul Kalam was interviewed by the Times of India
B: Dr.Abdul Kalam signed an order concerning mandatory education
C: Dr.Abdul Kalam opened opportunity in teaching profession
Related works, 1/22
D: mandatory Education foes' opinion
Related works, 1/26-1/29
E: comments on president's first week in office including his opinion on education

Fig 1. The study of the TDT4 topic "president Dr. Abdul Kalam insists on mandatory education" (1/13/2005 ~1/29/2005)

related paragraphs would be retrieved based on the subtopic. This is done my breaking the documents into smaller chunks and by tracking the word usage, subtopic boundaries are identified. But the limitation of this method is there is lack of information in the blocks in order to find the relationship between the other chunks of data. Brants et al. [2] and Choi et al. [7] made use of latent semantics believing that they would provide more information this model documents are treated as series of words and the documents are designed as states of HMM. They need data concerning the chunks. Blen and Moreno [17] made use of Hidden Markov Models to find the subtopic boundaries and in this model documents are treated as series of words and the documents are designed as states of HMM.

Topic identification and segmentation is a different process when compared to that of text segmentation by some means. Firstly, in document subtopic identification method single document is used where as with topic segmentation the input is set of related documents to a topic is used. Secondly, the identified segments of subtopic identification would be textual

paragraphs where as with topic segmentation; the identified actions of the themes have temporal property.

B. Text Summarization

Generic summarization focuses on covering more content area [24]. Here in our paper we focus on extraction based summarization where from the documents provided, the informative summaries are extracted. Exaction based methods can be of supervised or unsupervised forms. Shen et al. [1] proposed the method called conditional random fields (CRF's) to find the informativeness of a document. The top ranked sentences are collected as the summaries. Here large training corpora brought out a better result than the small training corpora. But only with suitable training corpora supervised summarization outperforms the unsupervised summarization. However there is a loop hole with supervised summarization, as they are domain dependency. Trying to use supervised summarization in new domain will result in more time consumption [22], [8]. Allan et al's summarizing method [10] summarizes the topics in a chronological order. In recent years graph- based summarization methods have caught the researcher's interest [8], [21], [3].zha [8] has made use of barpartite graph which when connected shows that the terms are more informative and this information score is upgraded iteratively. Mihalcea and Tarau's method [20] connects the similar pair of sentences. Then the summaries are obtained by using link analysis algorithms like HITS [20] and PageRank. Sun et al's method [24] used the queries that were used by the internet users to retrieve the information. These queries are based on the information that is present in the document. Nenkova et all [4] through his experimental results better than the summarization methods. proved that a simple term- frequency achieved its performance better than the summarization methods.

Topic summarization is a step higher and different from the above discussed text summarization as text summarization focus on a static content of document where as topic summarization method focus on temporal properties of the document.

C. Evolution of Topic

Graphical representations of topics are focused here. Kleinberg [26] proposed a topic evolution technique which used series of document to construct a hierarchical tree structure using HMM-based, two state Transition Diagram which in return helps in finding diverse themes from the different topics. Yang and Shi [14] proposed methods that focus on temporal properties of the documents. Feng and Allan [19] introduced incident threading method similar to our proposed system here incidents are derived from the document and these incidents are joined together to form the incident network. Swan and Allan [18] introduced a timeline system that displays significant topics graphically. We focus only on single topic and document that in return focus on documents that are related to that topic. In the Fig. 2 the system work flow diagram of the content summarization has been shown.

III. A CONTENT SUMMARIZATION SYSTEM

In this section, we portrait our methods used in the proposed content summarization system

A. Topic identification

A topic is that which comprises of one or more themes, which are related to certain issue. We define an action to be a part of theme that is very important to be taken into note by the readers. Basically all these actions together form the storyline(s) of the topic. Even though the actions do not seem to be interrelated temporarily they are considered to be semantically related since when they are joined together develops a theme. Fig.3 shows the relationship between themes, actions and action dependencies of a topic in the proposed system. Any topic is the collection of documents which are arranged in a chronological order. Our content summarization system will break the documents into non- overlapping segments. And a segment can be a numerous consecutive sentences. A segment or a block is defined as w consecutive sentences. Let the topic be represented by T and let T=$\{t_1, t_{2,},,, t_m\}$ be the set of vocabulary words without stop words [23]. And the topic can be represented as m x n term-segment association matrix B in which the columns are represented as $\{b_1, b_{2,},,, b_m\}$. They represent the blocks that are broken in a chronological order. Any weight of the term i in block j is performed using TF-IDF weighting scheme.

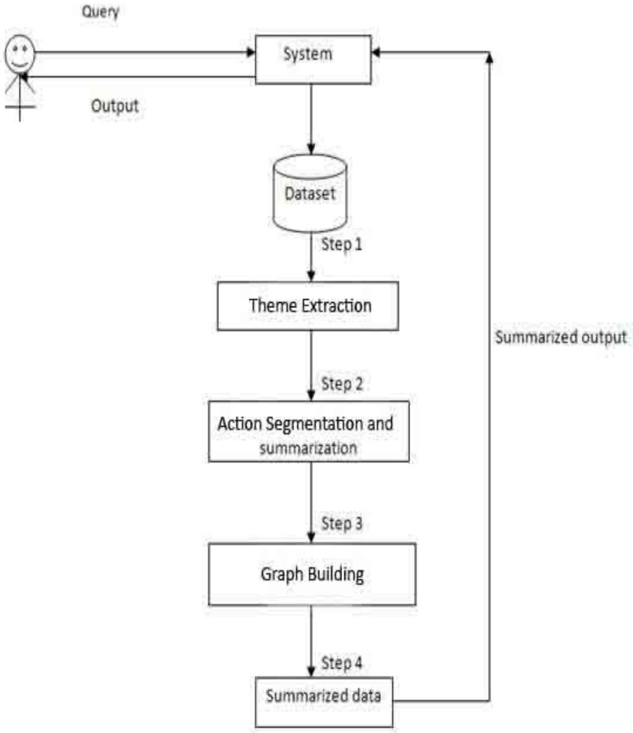

Fig 2. System Work Flow

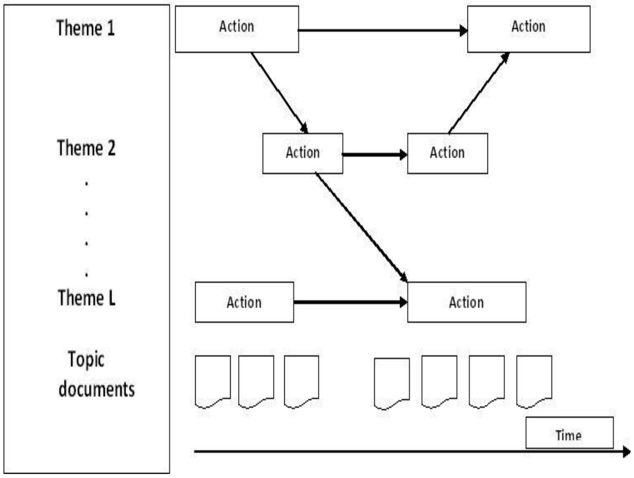

Fig 3.The Relationship between theme action & action dependency

B.Theme Extraction

The significant actions from the topic are called as themes. Themes can be generated through a segment association matrix. A matrix $A = B^T B$, called a segment association matrix, where A is an n x n symmetric matrix. The matrix entries with the larger values indicated that there is a high coherence between the pair of segments or blocks. Theme can be represented by an vector notation $\underline{v}$ of dimension n. Each vector entry shows how close the block is related to the theme. Given a vector $\underline{v}$, themes association with the topic is calculated using $\underline{v^T}$ A$\underline{v.}$ The show that the acquired theme is closely related to the theme, we adhere by the following function (1)

$$\max \underline{v^T}\ A\underline{v} \qquad (1)$$

$$s.t.\underline{v^T}\ \underline{v} = 1 \qquad (2)$$

The set of normalized vectors should be constrained to search within a search space else the value of v would be arbitrarily large hence the limitation is given in the function (2). The above two equations can be solved by the Legrangian formula [16].

$$Z(\underline{v},\lambda) = \underline{v^T}\ A\underline{v} + \lambda\ (1 - \underline{v^T}\ \underline{v}) \qquad (3)$$

To obtain the entry values of v, let ∂Z/∂v =∂Z/∂λ= 0 as follows

$$\partial Z/\partial \underline{v}=2A\underline{v}-2\lambda\underline{v}=0 \qquad (4)$$

$$\partial Z/\partial\lambda=1-\underline{v}^T\underline{v}=0 \qquad (5)$$

From equation (4) it is clear that Av = λv where v is a normalized eigenvector of A and λ is the corresponding eigenvalue. To obtain the relevant themes following theorem

Theorem 1

For any n x n symmetric matrix A of rank r, there exists a diagonal matrix D and an orthonormal basis V for R^n such that A = VDV^{-1}, where V= $\underline{v_1}$; $\underline{v_2}$; . . . ; $\underline{v_n}$ consists of the eigenvectors of A; and the diagonal entries of D satisfy $d_{1,1} \geq d_{2,2} \geq d_{r,r} > d_{r+1,r+1}$ = …. =$d_{n,n}$ which are eigenvalues corresponding to the respective columns of V.

Proof

V is the orthonormal basis of R^n and hence its inverse is identical to its transposition, i.e., $V^{-1} = V^T$[13]. The matrix representation of A is represented as below,

$$A = VDV^{-1} = VDV^T$$

$$= \lfloor \underline{v_1}, \ldots\ldots, \underline{v_n} \rfloor \lfloor d_1, 1\underline{e_1}, \ldots\ldots, d_{r,r\underline{e_r,}} 0_{e_{r+1}}, \ldots\ldots, 0_{\underline{e_n}} \rfloor V^T$$

$$= d_{1,1\underline{v_1}\underline{v_1^T}} + \ldots\ldots\ldots\ldots + d_{r,r\underline{v_r}\underline{v_r^T}} + \cdots\ldots\ldots 0_{\underline{v_n}\underline{v_n^T}} \qquad (6)$$

Where e_i denotes the standard vectors of R^n[13]. The themes of the topic can be considered by choosing first L significant eigenvectors where L<r. Then the interblock association of the selected themes can be approximated using following representation.

$$A \approx d_{1,1\underline{v_1}\underline{v_1^T}} + d_{2,2\underline{v_2}\underline{v_2^T}} + \ldots\ldots\ldots.d_{L,L\underline{v_L}\underline{v_L^T}}$$

$$= V_L D_L V_L^T \qquad (7)$$

WhereV_L, called theme matrix and $D_L$is a diagonal matrix whose diagonal entries are the top L eigenvalues of A. In order to have good coverage over the provided document, the eigenvectors of A are preferred to be orthogonal to each other.

C. Action Segmentation and Summarization

The entry $v_{i,j}$ points out the corresponding value of block or segment i and a theme j. And the theme is represented as $\underline{v_j}$ which is the normalized eigenvector and if the corresponding values are of high values then they are considered to be the actions that are found in the themes. If the values are smaller than they are the action boundaries. Kleinberg [20] and Nicholas and Dahlberg [12] showed that values of eigenvectors may be either positive or negative and it shows that certain information is found embedded in the document repository. In our summarization method first the actions are chunked down which is called as segmentation process and then they are summarized. Different types of summarization are performed as discussed in [24], [22], [8], [10], [21], [4] in order to cover various themes.

D .Graph Building

Themes and actions are obtained by above previous methods and in Graph Building phase, they connect the relevant themes and actions. Let there be a set X which contains the actions of the topic and they are represented as X={e_1, e_2,,, . . . , e_x}. Let the theme index of the actions be represented by $e_{k.}ev\in$ [1, L]. The action timestamp is represented by < e_k.fb; e_k.lb > where fb represents the index of the first block and lb represents the index of the last block of the actions. A directed acyclic graph is represented by G=(X,E) where X represents the set of action nodes and E represents the edges that join the different action nodes. Our approach which is action dependent [5], [11], [14], [19] involves two steps. First, the actions that are segregated from the same theme will be linked. Secondly, we will find the similarity between different actions segregated from different themes using temporal similarity (TS) and temporal weight (TW) functions and finally links them to form the storylines of the summaries. The Temporal Similarity function is calculated as below

$$TS(e_i, e_j) = TW(e_i, e_j) * cosine(e_i.cv, e_j.cv) \qquad (8)$$

Where, the cosine function returns the action centroid vector's cosine similarities. The temporal weight is calculated using the below equation,

$$TW(e_i, e_j) = \begin{cases} 1 - \frac{e_j.bb - e_i.eb}{n}, & if\ e_j.bb > e_i.eb \\ 1 - \frac{2*(min(e_i.eb, e_j.eb) - e_j.bb)}{|e_i| + |e_j|}, & if\ e_j.bb \leq e_i.eb \end{cases}$$

(9)

If the temporal similarity values is above the predefined threshold value then there construct a link between the two actions between which the temporal similarity is calculated.

IV.PERFORMANCE EVALUATIONS

A. Data corpus

In [26], two case studies using the official TDT topics proves that graph construction by content summarization system can refine and bring out themes, actions and action dependencies of the verified topics successfully. In this research we evaluate our summarization technique with that of several other text summarization techniques. Here we make use of official TDT4 topics whose corpora consists of 28,600 English news documents from five well known news agencies for the period of 1 January 2009 to 31 December 2009. Among them, 80 news actions with 2,056 related documents were labeled by NIST annotators for various tasks. We opted for 30 TDT4 topics, each containing more than 20 documents for each evaluation. Table 1 shows the detailed evaluated topics in our data corpus.

In the preprocessing phase, the document of the topic is segmented into chunks by using a Perl script provided by Document Understanding Conferences (DUC). H and w are the system parameters and to know how much they influence the summarization performance they are set at {5, 7, 9} and {1, 3, 5}, respectively. The parameter L is significant in finding the quality of the themes detected. The function U(L), defined below in (11), is used to find the underestimation values of V_L.

$$U(L) = \frac{1}{n*n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(a_{i,j} - (V_L D_L V_L^T)_{i,j}\right)^2 \qquad (12)$$

TABLE 1 STATISTICS OF EVALUATED TOPICS

Number of topics	30
Number of News documents	1,358
Average number of documents per topic	45.3
Number of sentences	34,547
Average number of sentences per topic	1,151.5

U(L) is the average of the squared differences between A and $V_L D_L V_L^T$. The value of U(L) is recommended to be low so that interblock association is sufficiently good in the contrary if they are high, then graph building phase might have too many themes to be comprehended. For summarization comparison, the evaluation are performed with L=1 to 10 in order to explain the influence of themes on the summarization performance. Fig. 4 shows the average U(L) of the 30 topics for L=1 to 10. It is observed that the segments with less content information produce a low underestimate. This is because segment association matrix is very sparse when they small in size. Hence, the value of U(L) is reduced when there is a slight difference between A and $V_L D_L V_L^T$.

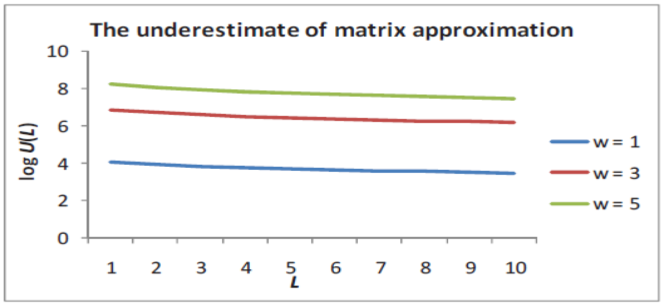

Fig 4.Underestimation of L themes

B. Scalability and Time comparisons

The execution time of summaries through different methods was evaluated on an AMD AthlonTM 64 Processor 3200++ PC with the Windows XP Service Pack 3 operating system and a 2 GB main memory. With a specified parameter setting the time taken to generate summaries was recorded. K – Means is a linear algorithm. FCW ia an iteration method in which for each iteration it compute the weight of every block based on a topic model; and its time complexity is O(n). TS method checks the content of the previous blocks to compute the blocks' novelty of the current block. And therefore its time complexity is O (n^2). Using MATLAB the eigenvectors of the matrix is computed and the complexity is found to be O ($n^2 I^2$) where I, is the number of eigenvectors to be calculated. The result in Fig. 5 shows that K-means and FCW methods run at greater speed than other methods. The execution time of our method increases as the summary size (L) increases. This is because for large values of L, the methods need to examine a lot of eigenvectors to compile summaries thus the execution time increases. For the K- means and FCW method, the number of cluster increases as L increases. Hence the method execution time also increases. It is found that TS methods execution time is irrelevant to the size of the summary because here all the topic blocks will be weighted irrespective of how many summary blocks are required.

The scalability of Content Summarization system for 30 topics is shown in Fig. 6. The parameters w, h and L are set at 1, 7 and 10, respectively. As the figure shows, for most of the topics it took around 20 seconds to compile topic summaries. For the largest topic which consists of 130 topic documents and 4,634 topic sentences, Content Summarization system approximately takes 2 minutes to construct the topic summary. However in real world this time is less than what any reader takes to read through these many documents individually. Thus they are accepted in real world implementation.

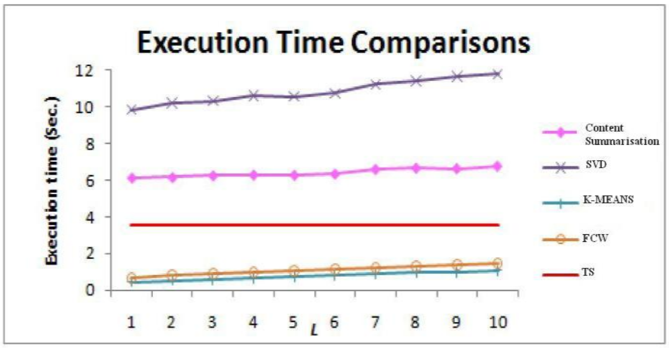

Fig 6. Comparison of execution times.

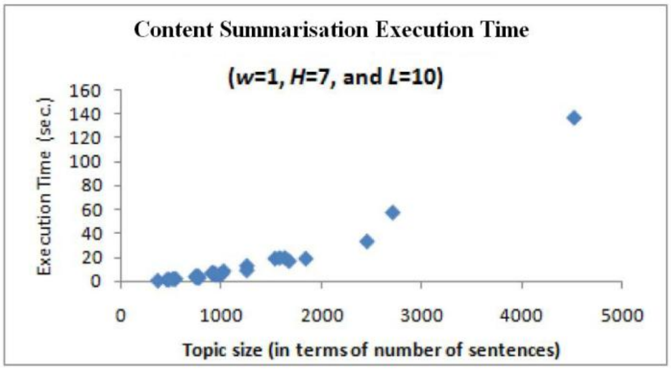

Fig 6. The scalability of Content Summarization.

V.CONCLUSIONS

It is quite common to publish news in the internet and any autonomous users can suggest their views on the published news. In order to get the glimpse of the document, content summarization system has been proposed. Summarization methods focus on covering wider area of content which comprises of important information. However, temporal properties of the information should be taken into consideration.

In this paper, we have discussed about Content Summarization system which extracts themes, actions and action summaries and how they are related to form the storyline(s) of the summary through graphical representation. TDT4 based experiments show that our Content Summarization system brings highly representative summaries.

ACKNOWLEDMENT

The authors would like to thank all the professors and lecturers who helped in completing this paper as well we thank the reviewers for their valuable comments and suggestions

References

[1] D. Shen, J.T. Sun, H. Li, Q. Yang, and Z. Chen, "Document Summarization Using Conditional Random Fields," Proc. 20th Int'l Joint Conf. Artificial Intelligence (IJCAI '07), pp. 2862-2867, 2007.

[2] T. Brants, F. Chen, and I. Tsochantaridis, "Topic-Based Document Segmentation with Probabilistic Latent Semantic Analysis," Proc. 11th Int'l Conf. Information and Knowledge Management, pp. 211-218,2002.

[3] R. Mihalcea and P. Tarau, "A Language Independent Algorithm for Single and Multiple Document Summarization," Proc. Int'l Joint Conf. Natural Language Processing, pp. 19-24, 2005.

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

[5] Q. Mei and C.X. Zhai, "Discovering Evolutionary Theme Patterns from Text—An Exploration of Temporal Text Mining," Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining,pp. 198-207, 2005.

[6] 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.

[7] F.Y.Y. Choi, P. Wiemer-Hastings, and J. Moore, "Latent Semantic Analysis for Text Segmentation," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 109-117, 2001.

[8] H. Zha, "Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering," Proc. 25th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 113-120, 2002.

[9] S. Strassel and M. Glenn, "Creating the Annotated TDT4 Y2003 Evaluation Corpus," http://www.itl.nist.gov/iad/mig/tests/ tdt/2003/papers/ldc.ppt, 2003.

[10] J. Allan, R. Gupta, and V. Khandelwal, "Temporal Summaries of News Topic," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 10-18, 2001.

[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. 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.

[13] L E. Spence, A.J. Insel, and S.H. Friedberg, Elementary Linear Algebra, a Matrix Approach. Prentice Hall, 2000.

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

[15] M.A. Hearst and C. Plaunt, "Subtopic Structuring for Full-Length Document Access," Proc. 16th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 59-68, 1993.

[16] W.L. Winston, Operations Research. Thomson, 2004.

[17] 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.

[18] 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.

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

[20] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment,"Proc. Ninth Ann. ACM-SIAM Symp. Discrete Algorithms, pp. 668-677, 1998.

[21] 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.

[22] 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.

[23] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.

[24] 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.

[25] X. Ji and H. Zha, "Domain-Independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming," Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 322-329, 2003.

[26] C.C. Chen and M.C. Chen, "TSCAN: A Novel Method for Topic Summarization and Content Anatomy," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 579-586, 2008.