A Novel Anatomy Approach For Document Summarization

Shahnaz Khadeeja M.
Dept.Of Computer Science and Engineering
M.E.A. Engineering College
Perinthalmanna, Kerala

Afsar P.
Assistant Professor
Dept. Of Computer Science and Engineering
M.E.A Engineering College
Perinthalmanna, Kerala

***Abstract*—The phenomenal growth in the number of documents posted on the Internet provides an abundant source of information as an alternative to traditional media. While current technologies are efficient in searching for appropriate documents to satisfy keyword search requests, users still have difficulty assimilating needed knowledge from the over- whelming number of documents. Hence an efficient summarization technique will be needed .In the proposed model we define task called topic anatomy, which summarizes and associates the core parts of topic temporarily 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 and then we can save our summary into a report format.**

Index Terms-Topic anatomy, Eigen vector,Context similarity , Association matrix, TSCAN.

I. INTRODUCTION

The World Wide Web has brought us a vast amount of on-line information. Due to this fact, every time someone searches something on the Internet, the response obtained is lots of different Web pages with many information, which is impossible for a personto read completely. So a summarization technique is essential in such cases. A textsummarizer strives to produce a condensed representation of its input, intended forhuman consumption. It may condense individual documents or groups of documents.Text compression, a related area, also condenses documents, but summarization differsin that its output is intended to be human-readable.

In this paper ,we have proposed the TSCAN approach for multi document summarization in web and resulting summarization is printed in report format. This method will help users to get required information in a report format with less effort, that means the user may not have to refer more documents. Instead of that a single keyword search will give the required summarized document .

TSCAN (Topic Summarization and Content ANatomy),[1] is a topic anatomy system which organizes and summarizes a temporal topic described by a set of documents. TSCAN consist of :

(a) Decomposing documents related to a topic into a non-overlapping sequence of blocks and describing a theme-identifying problem with blocks through a constraint optimization method to find and express themes as eigenvectors of a matrix;

(b) Analyzing changes in the eigenvectors through an R-S endpoint detection algorithm to detect events of each theme and obtain summarizations of the events

(c) Calculating context similarities of all of the events to obtain a temporal closeness per two events, and so forth, to form an evolution graph of the topic by associating all eventsaccording to the temporal closeness

.

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. In existing system, forward method, backward method, SVD method, K-means method, Temporal summary (TS) method, frequent content word method (FCW), TSCAN algorithm has been presented. These algorithms are used to find the close content for discovering text. The main problem in text mining is finding the closed pattern. These techniques are used for summarize the content.

In forward method, summarization is done by using initial block of content.[2] In this method, it will consider only initial block of text. This is main drawback of this method.

In backward method, summarization is done by using end block of content. In this method, it will consider only end block of text. This is main drawback of this method[2].

The SVD 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.[4] 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 eigen values of A'A.[10] 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 accessto 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.

The k-means algorithm is used for efficiency in clustering large data sets[5]. However, working only on numeric values prohibits it from being used to cluster real world data containing categorical values. The k-means algorithm 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 isan algorithm to partition and classify the data based on attributes or features in tok-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 .
3. It works only on numeric values. The K-means method which compiles summariesby 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 faircomparison of the K- means method, which provide the best result from 50 randomlyselected initial clusters for evaluation.

Temporal summary method is one of the summarization methods for content discovery. The temporal summary (TS) method take on the useful to and novel1 techniques proposed by the authors to compute the in formativeness score of a topic block.[6] 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.

Frequent content method 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[3]. In addition, we adopt Nenkova et al.'s context adjustment technique to increase the summary diversity

TSCAN method stands for Topic Summarization and Content Anatomy (TSCAN),[7]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, inwhich each block is a portion of a document, and treats each eigenvector of the matrixas a theme embedded in the topic. The eigenvectors are then examined to extract eventsand their summaries from each theme.The eigenvector are used for calculate the probability for extracting the content. Then,temporal similarity (TS) function is applied to generate the event dependencies, whichare then used to construct the evolution graph of the topic. Moreover, they are moreconsistent with human composed summaries than those derived by other text summarization methodsthat involves three major tasks: theme generation, event segmentationand summarization, and evolution graph construction.

TSCAN method are used to help the internet users graph grasp the gist of a topic coveredby a large number of topic documents, text summarization methods have been proposedto highlight the core information in the documents. Most summarization methods try toincrease the diversity of summaries to cover all the important information in the originaldocuments.

IV .PROPOSED WORK

Proposed anatomy approach for document summarization is a topic anatomy system which organizes and summarizes a temporal topic described by a set of documents. The first phase handles the crawling of the web pages to get text documents .This process is carried out according to the user query. The latter phase is segmentation and summarization process .This task is carried out in three phases. The documents are converted in to a matrix format , from that themes are generated followed by events. In the final phase the summary of the documents are generated and they are represented by a report format.

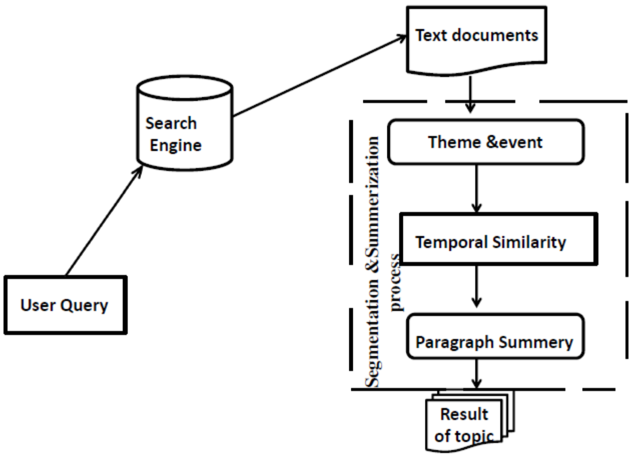

Figure 1.Proposed Architecture

A. System Modules

A detailed description on the implementation aspects of the proposed work are discussedin this section.

a. Crawling

We use the Web Crawler for retrieve the document. In a web page we retrieve the contents by the way of web crawler. We get the web page 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. Existing Google indexing is used for extracting documents.

b. Extraction

In this stage blocks are extracted from inbound links in the web sites. These sequence of non-overlapping blocks are decomposed from the crawled web documents. A block can be several consecutive sentences, or one or more paragraphs. Let $(b_1, b_2, \ldots, b_n)$represent the blocks chronologically decomposed from the topic documents, $T = (t_1, t_2, \ldots, t_m)$ be the set of stemmed vocabulary without stopwords of the topic. The topic can then be described by an mxn term-block association matrix B in which the row represent term and column represent blocks. For any two blocks bi and b_j, if i<j, then either the document containing bi was published before the document containing bj, or bi appears before bj in the same document. The (i, j)-entry of B (denoted as $b_{i,j}$) is theweight of term i in block j, computed by using the well known TF-IDF term weighting scheme

.

c. TF-IDF Term Weighting Scheme

In Vector model we de_ne weights using TF-IDF scheme[8]

1. wi;j is the weight for term ki in document dj :
$W_{i,j} = f_{i;j} * idf_i$

2. Normalized term frequency:
$f_{i,j} = freq_{i,j}=maxlfreq_{i,j}$
where the maximum is computed over all terms which are mentioned in the text of thedocument dj .
3. Term frequency $freq_{i,j}$ (ie, how often does term ki occur in document dj)
4. Inverse document frequency idfi : log(N=ni)
N = total number of documents
ni = number of documents in which ki occurs

d. Theme Generation

In this module we are identifying themes from the blocks that we are extracted. The identified themes are represented in terms of eigenvectors of a matrix. Next procedure is to generate events from these themes using R-S end point detection algorithm and hence to obtain the summary of the events. The steps included are given below :

1. Create a block association matrix A,
   $A = B^T .B$

B is the m*n matrix that we have already created. A is an nxn symmetric matrix in which the (i, j)-entry (denoted as $a_{i,j}$) is the inner product of columns i and j of the matrix B. Hence the matrix A represents the inter block association of the documents.

2. Next step is the theme generation:

A theme of a topic is regarded as an aggregated semantic profile of a collection of blocks. Theme can be represented as a vector v of 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.

3. The matrix A can be represented as follows:

$$A = VDV^{-1} = VDV^T$$
$$= [v_1,\ldots,v_n][d_{1,1}e_1,\ldots d_{r,r}e_r,0er+1,\ldots,0en]V^T$$
$$= [d_{1,1}v_1,\ldots d_{r,r}v_r,0v_r+1,\ldots,0v_n][v1,\ldots,vn]^T$$
$$= d_{1,1}v_1v_1^T+\ldots+ d_{r,r}v_rv_r^T+ 0v_{r+1}v_{r+1}^T+\ldots+ 0v_nv_n^T$$

Where $e_i$denotes the standard vectors of R_n *and* vector v_i, is an eigenvector of matrix A and $d_{i,i}$ is its corresponding eigen value 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:

$A \approx d_{1,1}v_1v_1^T + d_{2,2}v_2v_2^T + \ldots + d_{L,L}v_Lv_L^T$
$= [v_1, v_2, \ldots, v_L][d_{1,1}e_1,\ldots,d_{L,L}e_L][v_1, v_2, \ldots, v_L]^T$
$= V_LD_LV_L^T$

where V_L, called theme matrix, is an nxLmatrix in which a column represents a theme; and DL is an LxLdiagonal matrix in which the diagonal entries are the top L eigenvalues of A. i.e. the inter block association of a topic can be approximated by selecting a certain number of themes with significant eigenvalues. As the eigenvectors of Aare orthogonal to each other, the produced themes tend to be unique and descriptive.

e. Event Segmentationand Summarization

We adopt the R-S endpoint detection algorithm for event segmentation.(because: A theme vj in VL is a normalized eigenvector of dimension n, where the (i; j)-entry $v_{i,j}$ indicates the correlation between a block i and a theme j. As topic blocks are indexed chronologically, a sequence of entries in vj with high values can be considered as a noteworthy event embedded in the theme, and sequence of small values may be event boundaries. An eigenvector exhibit meaningful semantics for describing a certain concept embedded in a document corpus. To segment events, the endpoint detection algorithm examines the amplitude variation of a eigenvector to find the endpoints that partition the theme into a set of significant events.

In the R-S algorithm, every block in an eigenvector has an energy value, which is defined as follows:

$$eng_{(i,j)} = \frac{1}{H}\sum_{-H-1/2}^{H-1/2}\left[v_i + h_j\right]^2$$

whereeng$_{(i,j)}$ 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. A peak in the energy contour indicates that the corresponding sequence of blocks is a significant development of the theme and it is identified as an event.[9] To segment events from energy contours, we define a segmentation threshold as follows:

thd_{seg} = C _ maxi = 1….. n ; j = 1……L[eng(i; j)]

where C is in the range [0, 1], which is set as 0.2 in this study.[11] Then, linearly scan the energy contours for consecutive blocks whose energy values are above the threshold.For each event, the block with the largest amplitude is selected as the event summary.

f. Evolution Graph Construction

An evolution graph connects themes and events to present the storylines of a topic. Let X= {e1, e2, …,ex} be the set of events in a topic. For each event e_k, let $e_k.e_v \in$ [1, L] denote the theme index of the event, and <ek.bb, ek.eb> be the event's timestamp, where ek.bb and ek.ebare the indexes of the beginning and ending blocks, respectively. $|e_k|=1+e_k.e_b-e_k.bb$ is the temporal length of ek. The topic evolution graph G = (X, E) is a directed acyclic graph, where X represents the set of nodes and E= $\{(e_i, e_j)\}$ is the set of directed edges. An edge (e_i, e_j) specifies that event j is a consequent event of event i, which satisfies the constraint $e_j.bb > e_i.bb$. Identifying event dependency involves two procedures. First, we sequentially link events segmented from the same theme to reflect the theme's development. Second, we use a temporal similarity function to capture the dependency of events from different themes. For two events, eiand ej, belonging to different themes, where ej.bb >ei.bb, calculate their temporal similarity (TS) as:

$TS(e_i, e_j) = TW(e_i, e_j) * \text{cosine}(e_i.cv, e_j.cv)$,

Where the cosine function returns the cosine similarity between the centroid vectors of the events. The temporal weight(TW) function, weights the cosine similarity based onthe temporal difference between the events. If the temporal similarity is above a pre-defined threshold, then construct a link the corresponding events. Temporal weight is obtained based on a temporal relationship, such as non-overlapping, partial overlappingor complete overlapping, of the two events.Hence the themes and events are arranged intheir respective temporal similarity.

III. RESULTS AND OBSERVATIONS

Document summaries are difficult to evaluate, because for most applications there arenumerous summaries that are of equally high quality. Simply rewording portions of thesummary, reordering the sentences, omitting dubiously important information, etc., allresult in minor variations that are still excellent summaries. The initial core of our summarization approach is sentence extraction, so we can compare the sentences that a method chooses to the set of sentences that is known to be a good summary. To the extent that an approach chooses the right" sentences, that approach is good when it veers wildly from the ideal set, the approach is in appropriate to the task. Our approach is similar in spirit to the sentence based evaluations listed above, but is modified significantly to take into account the time-based nature of our summaries.

A. Data Extraction

In this method we are manually calculated the statics of topics.The documents aresearched and indexed using Google searching and indexing method. Total number of topicsevaluated are 30. The number of documents extracted are 180.Average number of documents per topic are60.. For each topic, average number of themes per topic are 77 and also average number of events per topic are 503

Table 1.Statistics of Evaluated Topics

Number of topics	30
Number of documents extracted	180
Average number of document per topic	60.0
Average number of themes per topic	77
Average number of events per topic	503

.

B. Summary Evaluation

In this section we are comparing our proposed system with previous methods such as1)The forward method, which generates summaries by extracting the initial blocks of atopic. 2) The backward method, which extracts summaries from the endblocks of atopic 3) The SVD method which composes summaries by extracting the blocks with thelargest entry value in singular vectors.
The summarization evaluation procedure is as follows. For each L,(L is the number ofthemes generated by a topic) we first apply our proposed work to each topic to extracta set of blocks as the topic summary. To ensure that the comparison with the other methods is fair, we use the compared methods' algorithms, and then produce summariesof the same size (in terms of the number of blocks) as those generated by our method.The compression ratios for summaries of L produced by the compared methods areshown in Table 2.

Table 2.Average size and compression ratios of summaries.

L valve	Proposed work		Forward method		Backward method		SVD method	
	Sum	C.R	sum	C.R	Sum	C.R	sum	C.R
1	8	97 %	6	94.5 %	6.5	95%	7	96%
2	12.7	96 %	10. 3	92%	11. 5	93.6 %	14. 9	94.1%
3	16.5	95 %	12. 8	91%	13. 1	92.3 %	15	94.8%

IV. CONCLUSION AND FUTURE WORK

In this work , we have presented a topic anatomy system called TSCAN, which extracts themes, events, and event summaries from crawled web documents. This systemhelps to analyze documents related to a topic through an eigenvector-based algorithmfor generating a summary and an evolution graph of the topic. This method obtains atemporal topic summary having a good quality with a consideration of topic temporality. It provides faster way to select representative sentences, paragraphs or documentsfor a topic while a compression ratio of summary is higher. This system helps to obtainan evolution graph showing important events in the topic and indicating cause-resultrelationships between the events for reducing difficulty in understanding an evolution ofthe topic. Also it helps users to get their search results in a summarized report format.Currently TSCAN approach is only used for documents which are written in a seriousand accurate manner, it may be difficult to apply the proposed method to other unconstrained texts, such as blogs. The related works our technology include NLP basedTSCAN method. In future instead of online news document, other unconstrained texts,such as blogs can be summarized. Also NLP based approach with large datasets projectsare under process.

REFERENCES

[1] C. C. Chen and M. C. Chen, \Tscan: a novel method for topic summarization and content anatomy," in Proceedings of the 31st annual international ACM SIGIRconference on Research and development in information retrieval. ACM, 2008, pp.579{586.

[2] P. Sathyashree et al., \A survey on content anatomy approach to temporal topic summarization," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 1, no. 10, pp. pp{310, 2012.

[3] A. P. Divya and S. Leela, \The content summarization system," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),vol. 2, no. 2, pp. pp{628, 2013.

[4] G. Erkan and D. R. Radev, \Lexrank: Graph-based lexical centrality as salience in text summarization," J. Artif. Intell. Res.(JAIR), vol. 22, no. 1, pp. 457{479, 2004.

[5] T. Nomoto and Y. Matsumoto, \A new approach to unsupervised text summarization," in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.ACM, 2001, pp. 26{34.

[6]J. Allan, R. Gupta, and V. Khandelwal, \Temporal summaries of new topics," in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.ACM, 2001, pp. 10{18.

[7] C. C. Chen and M. C. Chen, \Tscan: A content anatomy approach to temporal topic summarization," Knowledge and Data Engineering, IEEE Transactions on,vol. 24, no. 1, pp. 170{183, 2012.

[8] R. Baeza-Yates, B. Ribeiro-Neto et al., Modern information retrieval. ACM press New York, 1999, vol. 463.24References 25

[9] L. R. Rabiner and M. R. Sambur, \An algorithm for determining the endpoints of isolated utterances," Bell System Technical Journal, vol. 54, no. 2, pp. 297{315,1975.

[10]Y. Gong and X. Liu, \Generic text summarization using relevance measure andlatent semantic analysis," in Proceedings of the 24th annual international ACMSIGIR conference on Research and development in information retrieval. ACM,2001, pp. 19{25.

[11] R. Swan and J. Allan, \Automatic generation of overview timelines," in Proceedingsof the 23rd annual international ACM SIGIR conference on Research and development in information retrieval.ACM, 2000, pp. 49{56.