TSCAN: A Novel Method for Topic Summarization and Content Anatomy

Chien Chin Chen
Department of Information Management
National Taiwan University
Taipei, Taiwan
+886-2-3366-1194
paton@im.ntu.edu.tw

Meng Chang Chen
Institute of Information Science
Academia, Sinica,
Taipei, Taiwan
+886-2-2788-3799 ext. 1802
mcc@iis.sinica.edu.tw

ABSTRACT

A topic is defined as a seminal event or activity along with all directly related events and activities. It is represented as a chronological sequence of documents by different authors published on the Internet. In this paper, we define a task called topic anatomy, which summarizes and associates core parts of a topic graphically 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. Then, the significant events of the themes and their summaries are extracted by examining the constitution of the eigenvectors. Finally, the extracted events are associated through their temporal closeness and context similarity to form the evolution graph of the topic. Experiments based on the official TDT4 corpus demonstrate that the generated evolution graphs comprehensibly describe the storylines of topics. Moreover, in terms of content coverage and consistency, the produced summaries are superior to those of other summarization methods based on human composed reference summaries.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining, I.2.7 [**Natural Language Processing**]: Text analysis.

General Terms

Algorithms, Performance, Design, Experimentation.

Keywords

Topic anatomy. topic summarization, temporal text mining.

1. INTRODUCTION

The explosive growth in the number of documents available on the Internet has provided an abundant source of information as an alternative to traditional media. While current technologies can efficiently search 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 worse if the needed 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. 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) [1] project. The project defines a topic as "*a seminal event or activity, along with all directly related events and activities.*" Its goal is to automatically detect topics and track related documents from several document streams, such as on-line news feeds. The TDT project has attracted great deal of attention due to the importance and practicability of the problem. While an effective TDT system can detect topics and track all their related documents [3][5][19], users cannot fully comprehend a topic unless they read many of the tracked documents. Hence, there is an urgent need for effective summarization methods to extract the core parts of detected topics, as well as graphic representation methods to depict the relationship of core parts. Applied together the two techniques, called *topic anatomy*, can present essential information about a topic in a structured way.

Topic anatomy is an emerging text mining research issue that involves three major tasks: *theme generation*, *event segmentation and summarization*, and *evolution graph construction*. Generally, the content of a topic is comprised of several simultaneous *themes*, each representing an episode of the topic [11]. The theme generation process tries to identify the themes of a topic from the related documents. Over the lifespan of a topic, the focus of the topic's content may shift from one theme to another to reflect the topic's development [11]. We define an *event* as a disjoint sub-episode of a theme. The event segmentation and summarization process extracts topic events and their summaries by analyzing the intension variation of themes over time. Semantically, events may be associated because they are temporally close or share similar contexts, e.g., they may refer to the same named entities. By connecting the associations, the constructed evolution graph reveals the storylines of the topic. In this paper, we present a topic anatomy system called *TSCAN* (Topic Summarization and Content ANatomy), which organizes and summarizes a temporal topic described by a set of documents. TSCAN models the documents of a topic as a symmetric block (which is a portion of a document) association matrix 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. Finally, a temporal similarity function is applied to generate the event dependencies, which are used to construct the evolution graph of the topic. The results of experiments on the official TDT4 corpus demonstrate that the evolution graph construction process successfully extracts the themes, events, and event dependencies of the examined topics. Furthermore, compared to other text

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR'08, July 20-24, 2008, Singapore.
Copyright 2008 ACM 978-1-60558-164-4/08/07...$5.00.

summarization methods, our summaries are highly representative and compare well with human composed summaries.

2. RELATED WORKS

2.1 Text Summarization

Text summarization automatically creates a condensed version of one or more documents that captures the gist of the documents. In this study, we focus on extraction-based generic text summarization [6]. As a document's content may contain many themes, generic summarization methods concentrate on extending the summary's diversity to provide wide coverage of the content [7]. Gong and Liu [7] applied Singular Value Decomposition (SVD) [16] to the term-sentence association matrix of a document for extraction-based generic summarization. The authors regard the decomposed singular vectors as the themes of the document and compose diverse summaries by selecting informative sentences from important themes. Nomoto and Matsumoto [14] proposed the *X*-means algorithm, which groups the sentences of a document into theme coherent clusters. The algorithm is a variation of the standard *K*-means algorithm that estimates the number of clusters, i.e., *K*, dynamically during the clustering process. For each cluster, the sentence with the most information is selected as the summary. In recent years, graph-based summarization methods have generated a great deal of interest [6][20]. For example, Zha [20] modeled the relationship between sentences and terms of a document as a bipartite graph. The model considers a sentence informative if it connects with many informative terms, and vice versa. Zha proposes a reinforcement procedure that updates the informative scores of the terms and sentences iteratively. Finally, summaries are composed by selecting informative sentences. Erkan and Radev [6] represented a set of documents as a graph in which the nodes stand for sentences and content-similar sentences are associated with edges. The assumption is that a sentence is informative if it connects with many sentences; hence, by extension, the connected sentences are also informative. By deriving the informative scores of sentences from their connected sentences iteratively, the most informative sentence can be taken as the summary.

Topic summarization differs from traditional text summarization because of its temporal properties. As topics are reported chronologically, comprehensive topic summaries should describe the evolution (i.e., storylines) of the topics in addition to possessing informative sentences.

2.2 Topic Evolution Mining

Kleinberg [9] developed the technique of topic evolution mining by constructing a hierarchical tree on a series of topic documents. He utilized a HMM-based two-state transition diagram to model the status of topics and suggested that a topic can be split into diverse themes, modeled as tree branches, whenever it receives bursty information. Nallapati et al. [12] formalized the problem of topic evolution mining as a text clustering task in which the identified clusters, i.e., the events of a topic, are connected chronologically to form an evolution graph of the topic. In addition to graph construction, Mei and Zhai [11] modeled the activeness trend of identified themes. As the trend reveals variations in the activeness of a theme over the lifespan of a topic, it helps users catch the rise and fall of the topic's development. Yang and Shi [18] focused on the temporal properties of a topic, and showed that fine-grained evolution graphs can be obtained by using the temporal information about topics.

3. TSCAN: A TOPIC ANATOMY SYSTEM

3.1 Topic Model

A topic is a real world incident that consists of one or more themes, which are related to a finer incident, a description, or a dialogue of a certain issue. During the lifespan of a topic, one theme may attract more attention than the others, and is thus reported by more documents. We define an event as a significant theme development that continues for a period of time. Although the events of a theme are temporally disjoint, they are considered semantically dependent to express the development of the theme. Moreover, events in different themes may be associated due to their temporal closeness and context similarity. Naturally, all the events taken together form the storylines of the topic. The proposed method identifies themes and events from the topic documents, and connects associated events to form the topic evolution graph. In addition, the identified events are summarized to help readers better comprehend the storylines of the topic. Figure 1 illustrates the relationship between the themes, events, and event dependencies of a topic.

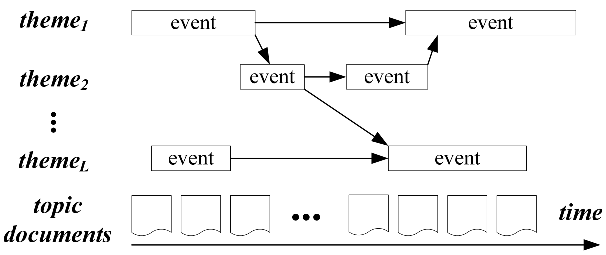

Figure 1. The relationships of themes, events, and event dependencies.

A topic is explicitly represented by a collection of chronologically ordered documents. In this study, we assume that the documents are published in the same order as the reported events of the topic, and that there is no inconsistency between the contents of documents. Each document in TSCAN is decomposed into a sequence of non-overlapping *blocks*. A block can be several consecutive sentences, or one or more paragraphs. We define a block as *w* consecutive sentences. Let $T=\{t_1, t_2, \ldots, t_m\}$ be the set of stemmed vocabulary [4] without stopwords [4] of the topic. The topic can then be described by an *mxn term-block association matrix B* in which the columns $\{\underline{b}_1, \underline{b}_2, \ldots, \underline{b}_n\}$ represent the blocks chronologically decomposed from the topic documents. In other words, for any two blocks $\underline{b}_i$ and $\underline{b}_j$, if $i<j$, then either the document containing $\underline{b}_i$ was published before the document containing $\underline{b}_j$, or $\underline{b}_i$ appears before $\underline{b}_j$ in the same document. The (i,j)-entry of B (denoted as $b_{i,j}$) is the weight of term i in block j, computed by using the well-known TF-IDF term weighting scheme [4].

3.2 Theme Generation

The matrix $A=B^TB$, called a *block association matrix*, is an *n*x*n* 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. As a column of B is the term vector of a block, A represents the inter-block association. Hence, 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, and can be represented as a vector $\underline{v}$ of dimension n, where each entry denotes the degree of correlation of a block to the theme. Given a constitution of a vector $\underline{v}$, $\underline{v}^TA\underline{v}$ computes the association of the theme to the topic's content. The objective function shown in Eq. (1)

of our theme generation process determines $\underline{v}$'s entry values so that the acquired theme is closely associated with the topic.

$$\max \underline{v}^T A \underline{v} \quad . \tag{1}$$

$$\text{s.t. } \underline{v}^T \underline{v} = 1. \tag{2}$$

Without specifying any constraint on $\underline{v}$, the objective function (1) becomes arbitrarily large with large entry values of $\underline{v}$. Constraint (2) limits the search space within the set of normalized vectors such that the following Lagrangian formula [17] can be used to solve Eqs. (1) and (2).

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

To obtain the entry values of $\underline{v}$, let $\partial Z/\partial \underline{v} = \partial Z/\partial \lambda = 0$ as follows:

$$\partial Z/\partial \underline{v} = 2A\underline{v} - 2\lambda \underline{v} = 0. \tag{4}$$

$$\partial Z/\partial \lambda = 1 - \underline{v}^T \underline{v} = 0. \tag{5}$$

Eq. (4) implies that $A\underline{v}=\lambda\underline{v}$. In other words, $\underline{v}$ is a normalized eigenvector of A and λ is the corresponding eigenvalue. For any nxn square matrix, there can be at most n eigenvectors [16]. In terms of non-linear programming, Eq. (3) can have more than one stationary point [17]. To acquire appropriate themes of the topic, the following theorem of symmetric matrices is employed.

Theorem 1. *For any nxn 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* $V=\{\underline{v}_1,\underline{v}_2,\ldots,\underline{v}_n\}$ *consists of the eigenvectors of A; and the diagonal entries of D satisfy* $d_{1,1} \geq d_{2,2} \geq \ldots \geq d_{r,r} > d_{r+1,r+1} = \ldots = d_{n,n}=0$, *which are eigenvalues corresponding to the respective columns of V.*

The proof of the theorem can be found in the section on diagonalization of symmetric matrices of many linear algebra books [16]. Since V is an orthonormal basis of R^n, its inverse is identical to its transposition, i.e., $V^{-1}=V^T$ [16]. Therefore, the matrix A can be represented as follows:

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

$$= [\underline{v}_1,\ldots,\underline{v}_n][d_{1,1}\underline{e}_1,\ldots d_{r,r}\underline{e}_r, 0\underline{e}_{r+1},\ldots,0\underline{e}_n]V^T$$

$$= [d_{1,1}\underline{v}_1,\ldots d_{r,r}\underline{v}_r, 0\underline{v}_{r+1},\ldots,0\underline{v}_n][\underline{v}_1,\ldots,\underline{v}_n]^T$$

$$= d_{1,1}\underline{v}_1\underline{v}_1^T + \ldots + d_{r,r}\underline{v}_r\underline{v}_r^T + 0\underline{v}_{r+1}\underline{v}_{r+1}^T + \ldots + 0\underline{v}_n\underline{v}_n^T, \tag{6}$$

where $\underline{e}_i$ denotes the standard vectors of R^n [16]. In other words, the symmetric matrix A can be decomposed into the sum of n matrices spanned by its eigenvectors. We treat the first L ($L<r$) significant eigenvectors of A as the themes of the topic. Then, the inter-block association approximated by the selected themes can be represented as follows:

$$A \approx d_{1,1}\underline{v}_1\underline{v}_1^T + d_{2,2}\underline{v}_2\underline{v}_2^T + \ldots + d_{L,L}\underline{v}_L\underline{v}_L^T$$

$$= [\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_L][d_{1,1}\underline{e}_1,\ldots,d_{L,L}\underline{e}_L][\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_L]^T$$

$$= V_L D_L V_L^T, \tag{7}$$

where V_L, called *theme matrix*, is an nxL matrix in which a column represents a theme; and D_L is an LxL diagonal matrix in which the diagonal entries are the top L eigenvalues of A. In short, the inter-block association of a topic can be approximated by selecting a certain number of themes with significant eigenvalues. Note that, as the eigenvectors of A are orthogonal to each other, the produced themes tend to be unique and descriptive.

3.3 Event Segmentation and Summarization

A theme $\underline{v}_j$ in V_L 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 $\underline{v}_j$ with high values can be considered as a noteworthy event embedded in the theme, and valleys (i.e., a sequence of small values) in the entry sequence may be event boundaries. However, according to the definition of eigenvectors [16], the sign of entries in an eigenvector is invertible. Moreover, [8] shows that both the positive and negative entries of an eigenvector exhibit meaningful semantics for describing a certain concept embedded in a document corpus. Therefore, we adopt the R-S endpoint detection algorithm [15], which examines the variation in the amplitude of an eigenvector, for event segmentation. Endpoint detection and event segmentation are similar in that they both try to find separation points between major segments of sequential data. To segment events, the endpoint detection algorithm examines the amplitude variation of an 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:

$$eng(i,j) = \frac{1}{H}\sum_{h=-(H-1)/2}^{(H-1)/2}[v_{i+h,j}]^2, \tag{8}$$

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 neighborhood.

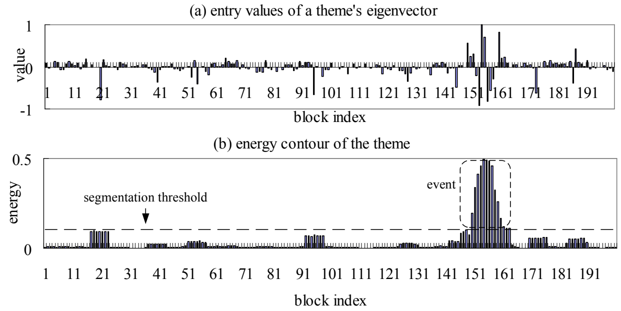

Figure 2. The eigenvector of a theme and its energy contour.

Figure 2 shows the eigenvector of a theme and its energy contour. A peak in the energy contour (e.g., from indexes 150 to 163) indicates that the corresponding sequence of blocks is a significant development of the theme; therefore, it is identified as an event. To segment events from energy contours, we define a segmentation threshold as follows:

$$thd_{seg} = C * \max_{i=1\ldots n, j=1\ldots L}[eng(i,j)], \tag{9}$$

where C is in the range [0,1], which is set as 0.2 in this study. Then, we linearly scan the energy contours for consecutive blocks whose energy values are above the threshold. To reduce false alarms about event segmentation and refine the segmentation result, we adopt the two frequently used heuristics: 1) two close events are merged, and 2) small events are eliminated [15]. For each event, the block with the largest amplitude is selected as the event summary. Note that the summary block might not be the one with the largest energy value due to the averaging effect of the sliding window. Another interesting by-product of the above method is that the produced energy contour also describes the activeness trend of a theme. In the experiment section, we demonstrate that the changes in energy contours accurately reflect the evolution of a theme.

A unique feature of our summarization approach is the introduction of the event segmentation process to extract the semantic construct “event” before summarization. Most existing generic summarization approaches [6][7][14][20] try to cover diverse themes in document summaries, but our method further describes the development of themes via summarized events to help users comprehend the storylines of a topic.

3.4 Evolution Graph Construction

An evolution graph connects themes and events to present the storylines of a topic. Let $X=\{e_1, e_2, \ldots, e_x\}$ be the set of events in a topic. For each event e_k, let $e_k.ev \in [1, L]$ denote the theme index of the event, and $<e_k.bb, e_k.eb>$ be the event's timestamp, where $e_k.bb$ and $e_k.eb$ are the indexes of the beginning and ending blocks, respectively; $|e_k|=1+e_k.eb–e_k.bb$ is the temporal length of e_k. 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 $e_j.bb > e_i.bb$.

Automatic induction of event dependency is often difficult due to the lack of sufficient domain knowledge and effective knowledge induction mechanisms [12]. However, previous works [11][12] [18] have shown that, instead of domain knowledge, word usage analysis can identify event dependency. Our approach follows this rationale and involves two procedures. First, we sequentially link events segmented from the same theme to reflect the theme's development. Then we use a temporal similarity function to capture the dependency of events from different themes. For two events, e_i and e_j, belonging to different themes, where $e_j.bb > e_i.bb$, we calculate their *temporal similarity* (*TS*) by Eq. (10):

$$TS(e_i,e_j) = TW(e_i,e_j) * cosine(e_i.cv,e_j.cv), \tag{10}$$

where the *cosine* function returns the cosine similarity [4] between the centroid vectors of the events. The centroid vector, $e_k.cv$, of an event e_k is defined as follows:

$$e_k.cv = \frac{1}{|e_k|}\sum_{i=e_k.bb}^{e_k.eb} |v_{i,e_k.ev}| * \underline{b}_i, \tag{11}$$

where $\underline{b}_i$ is the term vector of block i. In short, $e_k.cv$ averages the term vectors of the event blocks in accordance with their correlation to the event. The *temporal weight* (*TW*) function, defined in Eq. (12), then weights the cosine similarity based on the temporal difference between the events. If the temporal similarity is above a pre-defined threshold, we deem e_j a consequence of e_i and construct a link between them.

$$TW(e_i,e_j) = \{ \begin{matrix} 1-\frac{e_j.bb-e_i.eb}{n}, & \text{if } e_j.bb>e_i.eb, \\ 1-\frac{2*(\min(e_i.eb,e_j.eb)-e_j.bb)}{|e_i|+|e_j|}, & \text{if } e_j.bb\leq e_i.eb. \end{matrix} \tag{12}$$

The range of the proposed *TW* function is within (0,1]. As shown in Figure 3, *TW* considers the temporal relationship between events e_i and e_j and gives an appropriate temporal weighting. In case 1 where e_i and e_j do not overlap, *TW* penalizes the events with a large temporal distance.

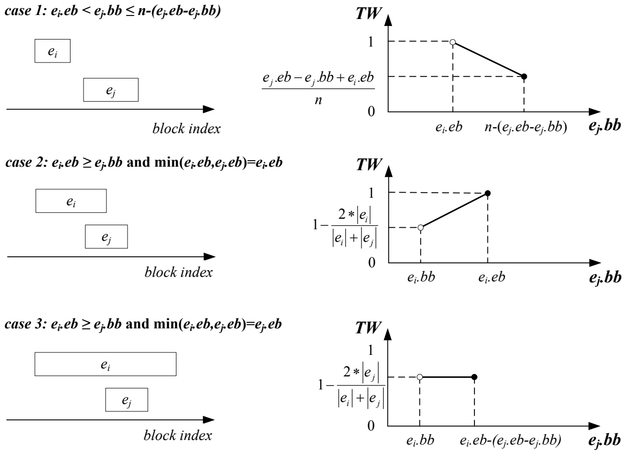

Figure 3. The graphs of the *TW* function under three temporal cases.

The penalty corresponds to Yang's observation [19] that temporally close pieces of information are usually more relevant to one another than those far apart. In case 2, e_i and e_j do overlap and e_j is not contained in e_i. *TW* penalizes events if their beginning timestamps are close to each other. This penalty is based on the supposition that when two events happen almost simultaneously, they are probably the distractions of a certain prior event, rather than being dependent on one another. For instance, the outcome of a baseball tournament may give rise to concurrent events of celebrations and player trades. In case 3, e_j is contained in e_i, and the value of *TW* decreases with the increase in $|e_j|$. This property, similar to case 2, prevents linking events with similar timestamps because they may be distractions of a prior event.

4. PERFORMANCE EVALUATIONS

In this section, we evaluate the performance of TSCAN. Traditionally, performance evaluations in information retrieval depend on annotated benchmarks. However, to the best of our knowledge, there are no official benchmarks and metrics for the study of topic anatomy. Therefore, in this section, 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 TDT4 corpus [1] where 26 news topics, each containing more than 20 documents, were selected for performance evaluations. The topics were labeled by NIST annotators and their elaborate explications are regarded as reference summaries for summarization evaluations. Although DUC (Document Understanding Conferences) [2] also use TDT topics for summarization contests, the average size of the DUC topics is only 10 documents, which is too small to demonstrate the effect of the proposed method.

In the pre-processing phase each topic document is partitioned into blocks of sentences by using a simple script supplied by DUC. To reduce the impact of sentence partition errors, each block has 3 sentences to ensure that it contains a complete sentence. The parameter L is critical to the quality of detected themes. From Eq. (6), it is clear that the larger the number of themes selected, the better the approximation will be. For summarization comparisons, the evaluations are performed with L=1 to 10 to show the influence of themes on summarization performance. In addition, the parameter H and the temporal similarity threshold are set to 7 and 0.3, respectively.

4.1 Summarization Evaluations

We compare the summarization performance of TSCAN with the following four summarization methods: 1) The forward method, which generates summaries by extracting the initial blocks of a topic. 2) The backward method, which extracts summaries from the end blocks of a topic. This is frequently used as the baseline method in DUC contests [13]. 3) The SVD method [7], which composes summaries by extracting the blocks with the largest entry value in singular vectors. Note that the result of the SVD method is identical to that of the graph-based summarization method [6]. 4) The *K*-means method [14], which composes summaries by selecting the most salient blocks of the resulting *K* clusters. Generally, the performance of the *K*-means method depends on the quality of the initial clusters. In this experiment, to give the *K*-means method fair consideration, the best result from fifty randomly selected initial clusters is used for comparison.

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. To ensure that the comparison with the other methods is fair, 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 are shown in Table 1. In sum, the compression ratios of the evaluated summaries are high and at least 90% of the topic's contents are omitted.

Table 1. Average size and compression ratios of summaries.

L	1	2	3	4	5	6	7	8	9	10
Sum.	8.7	12.5	16.5	20.7	23.8	27.8	30.5	33.8	37.2	41.2
C.R.	98%	97%	96%	95%	94%	93%	93%	92%	91%	90%

Legend: Sum.: summary size, the number of blocks in the topic summary; C.R.: compression ration, 1-summary size / topic size.

We use two metrics, summary-to-document content similarity (SDCS) and ROUGE [10], to evaluate the above summarization methods. The SDCS metric compares the content coverage of a generated summary to the documents for summarization, while ROUGE considers the consistency between the content of a generated summary and that of a set of expert-composed reference summaries.

4.1.1 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.

Table 2. Micro average summary-to-document content similarity.

	TSCAN	Forward		Backward		SVD		*K*-means	
L	SDCS	SDCS	IMP	SDCS	IMP	SDCS	IMP	SDCS	IMP
1	0.384	0.280	37%	0.221	73.8%	0.336	14.3%	0.355	8.2%
2	0.380	0.309	23%	0.244	55.6%	0.347	9.5%	0.367	3.6%
3	0.381	0.325	17.3%	0.267	43%	0.346	10.3%	0.382	-0.2%
4	0.385	0.337	14.2%	0.285	35.1%	0.348	10.7%	0.393	-2.2%
5	0.389	0.346	12.5%	0.299	30.1%	0.349	11.3%	0.401	-3.1%
6	0.388	0.357	8.5%	0.311	24.6%	0.355	9.3%	0.403	-3.8%
7	0.390	0.362	8%	0.320	22%	0.366	6.8%	0.413	-5.5%
8	0.390	0.366	6.8%	0.328	18.9%	0.371	5.2%	0.420	-7.1%
9	0.389	0.370	5.2%	0.337	15.5%	0.373	4.4%	0.424	-8.2%
10	0.395	0.377	4.6%	0.347	13.9%	0.369	6.8%	0.429	-8.1%

Legend: SDCS: the micro average summary-to-document content similarity; IMP: the improvement by our method.

As shown in Table 2, our method outperforms the compared methods, except the *K*-means method with large *L* values. The latter achieves a higher similarity score because its summary provides better coverage of the topic's contents. Our method simply selects *L* most significant themes (*L*<<*r*) to represent a topic, whereas the *K*-means method partitions the all of the topic's content into *K* clusters and extracts the most salient block from each cluster to represent the topic. As a result, summaries constructed by the *K*-means method provide better content coverage, and the similarity score increases as more clusters are used to partition the content. However, without an effective mechanism, such as the structure of themes and events, to leverage and organize the summarized results, large *K* values indicate that the summaries are unstructured, and therefore difficult for users to understand. Both the proposed method and the SVD method perform in between. Our method outperforms the SVD method, especially when the compression ratio is high. This is because our summary favors significant themes and events, which are representative of topic documents. The coverage provided by forward and backward methods is poor because their summaries only cover the beginning and end of topic documents, respectively. To improve their performance, lower compression ratios are required so that more topic information can be included in the summaries.

4.1.2 ROUGE Evaluations

ROUGE is a recall-oriented summary evaluation metric used mostly in DUC contests [13]. 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. It has been shown that the results of the comparisons based on ROUGE-1 (unigram overlapping) and ROUGE-2 (bi-gram overlapping) are consistent with manual judgments [10]. Therefore, we use ROUGE-1 and ROUGE-2 to evaluate the consistency of manual summaries derived by the compared methods. Tables 3 and 4 show the micro average performances of ROUGE-1 and ROUGE-2, respectively.

As ROUGE is a recall-oriented evaluation metric, the scores of all the compared methods increase with the increases in *L*. In [13] the backward method is regarded as effective because its ROUGE performances have proved comparable to those of many elaborate summarization methods in a number of contests. However, in our evaluations, the backward method is inferior to the simple forward method. This is because the first few sentences of a news article usually detail the essential part of the story. Thus, the forward method is more effective and, in fact, it is nearly comparable to the sophisticated SVD method. The results show that the proposed method achieves the best ROUGE-1 and ROUGE-2 scores for all *L*. Moreover, the improvement it achieves over the compared methods increases as the compression ratio increases (i.e., a decrease in *L*). For example, when *L*=1, our method outperforms the compared methods by 27.3% to 62.6% for ROUGE-1 and by 41.7% to 125% for ROUGE-2. This is because our method selects representative sentences earlier than the compared methods when composing topic summaries. In resource-limited environments, such as low network bandwidth or the small display panels of hand-held devices, this property helps users capture key information about a topic. The superior ROUGE performance of TSCAN is related to the manner of summary composition. The *K*-means method and the SVD method increase summary coverage by using clusters and singular vectors, respectively, whereas our method distinguishes between important events in themes to achieve both summary diversity and narrative tracing properties. The results of TSCAN are consistent with topic annotators' reference summaries, which generally explain the major events of significant themes. As a result, TSCAN outperforms the compared methods in terms of the ROUGE evaluation metric.

Table 3. The micro average performance of ROUGE-1 in the compared methods.

	TSCAN	Forward		Backward		SVD		*K*-means	
L	R-1	R-1	IMP	R-1	IMP	R-1	IMP	R-1	IMP
1	0.524	0.365	43.5%	0.322	62.6%	0.401	30.7%	0.412	27.3%
2	0.564	0.448	25.7%	0.401	40.6%	0.464	21.4%	0.475	18.7%
3	0.594	0.490	21.2%	0.452	31.3%	0.499	19%	0.523	13.4%
4	0.613	0.518	18.3%	0.484	26.7%	0.542	13%	0.555	10.5%
5	0.623	0.538	15.8%	0.502	24.1%	0.557	11.8%	0.581	7.2%
6	0.646	0.565	14.2%	0.537	20.2%	0.571	13%	0.601	7.4%
7	0.651	0.578	12.6%	0.557	16.9%	0.591	10.2%	0.615	5.8%
8	0.662	0.595	11.2%	0.578	14.6%	0.608	8.8%	0.646	2.4%
9	0.670	0.618	8.4%	0.598	12%	0.621	7.9%	0.651	2.9%
10	0.681	0.647	5.2%	0.626	8.7%	0.639	6.5%	0.671	1.5%

Legend: R-1: the micro average ROUGE-1 performance.

Table 4. The micro average performance of ROUGE-2 in the compared methods.

	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.

4.2 Evolution Graph Evaluations

Two TDT4 topics, #40023 "President Bush Bans Abortion Funding" and #40004 "Russian Nuclear Submarine Kursk Sinks" are selected as case studies of topic evolution graphs because their stories are well known and readers can understand them without specific knowledge or information about the cultural background.

4.2.1 Case Study 1 — Topic 40023

The topic "President Bush Bans Abortion Funding" relates to the president's decision on abortion in January, 2001. The topic contains 25 documents and 253 blocks. Figure 4 shows the constructed topic evolution graph, which consists of 5 themes (i.e., *L*=5), 13 identified events, and 14 edges. Each row of the graph shows the events of a theme. For each event, we show the indexes of its beginning, ending, and summary blocks as a trinary tuple.

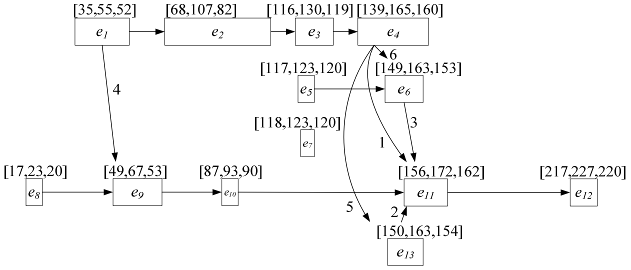

Figure 4. The evolution graph of Topic 40023.

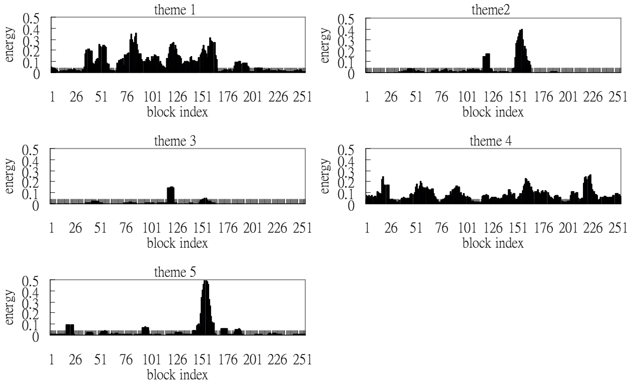

Figure 5. The energy contours of the themes in Topic 40023.

According to the summaries, Theme 1 relates to President Bush's attitude toward abortion. Theme 2 represents the opinions of conservative groups on abortion. Theme 3 discusses the case of "Roe versus Wade," which led the United States Supreme Court to legalize abortion in 1973. Theme 4 describes the interaction between President Bush and his cabinet. Finally, Theme 5 considers the impact of banning funding for abortion on medical research. Basically, these themes not only cover the storylines of the topic, but also include many interesting items of side information. For example, as the topic's time period was just the president's first week in office, the topic documents also include the interactions between him and the new cabinet. Surprisingly, the interactions were included in Theme 4. Furthermore, Theme 3 deals with the anniversary of the Supreme Court's decision on Roe versus Wade, which was on the president's first day in office.

On the generated evolution graph and in the summaries, the topic was initiated by event e_8 when president-elect Bush revealed during a media interview that he might reverse some of President Clinton's orders. Then, in event e_1, President Bush announced that he would issue an executive order banning funding for abortion. Soon afterwards, in event e_2, he signed a memorandum about blocking the funding. As the periods of e_1 and e_2 covered his first day in office, the topic-related documents also mentioned his goals, and his interaction with the new cabinet. These incidents were extracted by events e_9 and e_{10}. Coincidentally, the first day in office was also the anniversary of the Roe versus Wade decision, which was identified by event e_7 and the opinions of pro-life groups were the focus in event e_5. Next, in event e_4 Bush declared that support for stem cell research should be cut off because stem cells are taken from aborted fetuses. This declaration immediately attracted comments from conservative groups (e_6) and medical researchers (e_{13}). Bush reaffirmed that the government would not support research related to aborted fetuses (e_{11}). Finally, in event e_{12}, the media commented on the president's first week in office, including his decision to ban abortion. They also mentioned the difficulties and challenges that the president and his cabinet would face during his presidency.

The energy contours shown in Figure 5 also reflect the trends of the themes. The contours of Themes 1 and 4 have several peaks distributed uniformly over the time period of the topic. Since "Bush's attitude toward abortion" and "the interaction between Bush and his cabinet" are the main themes of the topic, nearly all the topic's documents are associated with a particular context. As a result, their energy distributions are uniform. Additionally, the peaks of Themes 3 and 5, which correspond to the time phases of the Roe versus Wade anniversary and the debate on medical research, respectively, demonstrate that the energy contours of eigenvectors can describe the trends of themes.

Finally, we examine the quality of links in the evolution graph. Generally, the links of events of the same theme state the storyline of the theme well. For example, the links of Theme 1 sequence the different phases of President Bush's attitude toward abortion from probability to certainty. Moreover, the links of different themes also explain the associations between events successfully. We label these kinds of links according to their ranking in terms of the *TS* values. As the criterion of link construction is based on the content similarity discounted with a (0,1] temporal weight, the linked events always have similar contexts and thus exhibit meaningful associations. For example, the set of links {1, 2, 3, 5, 6} clearly states that Bush's decision in event e_4 caused disagreements between researchers (e_{13}) and conservative groups (e_6), which forced him to reaffirm his determination in e_{11}. The style of links also demonstrates the utility of the proposed temporal weight function. As mentioned earlier, the temporal weight function prevents linking events that are separated by large temporal distances. Thus, the resulting evolution graph rarely has interlaced large-distance links, which makes the graph concise and traceable. Moreover, the function also distinguishes parallel events, such as e_5 and e_7, both of which have similar content with regard to Roe versus Wade, but neither one draws out the other. In sum, the proposed method successfully extracts meaningful events and summary sentences, and organizes their dependencies well so that users can easily comprehend the storylines of the topic.

4.2.2 Case Study 2 — Topic 40004

The second example is based on TDT4 40004 "Russian Nuclear Submarine Kursk Sinks" which reported the recovery of bodies from the sunken Russian submarine Kursk in October 2000. The topic is larger than the first example and consists of 329 blocks in 56 related documents. Figures 6 and 7 show the constructed evolution graph and energy contours, respectively.

According to the summaries, Theme 1 relates to the progress of body recovery. Theme 2 describes how rescue divers salvaged the sunken submarine. Theme 3 discusses when and how the sailors died. Theme 4 is related to legislation that prevented Russians from using words from foreign languages when speaking Russian. Finally, Theme 5 discusses the causes of the explosion and sinking of the submarine. In fact, Theme 4 has nothing to do with the topic. However, the documents of Theme 4 mentioned that Russian President Vladimir Putin was the target of the legislative bill because he used inappropriate foreign words when speaking to the families of the sailors killed in the submarine disaster. This is probably the reason that TDT annotators included the documents in the topic. Nevertheless, it is interesting to note that the proposed method can detect this incident and treats it as an isolated event (i.e., e_{14}) in the evolution graph.

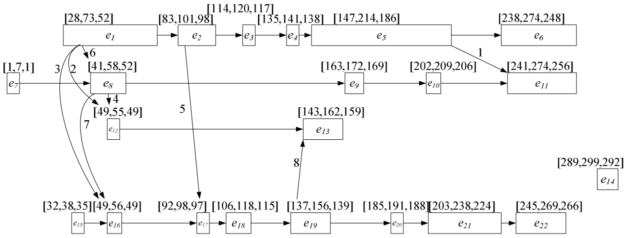

Figure 6. The evolution graph of Topic 40004.

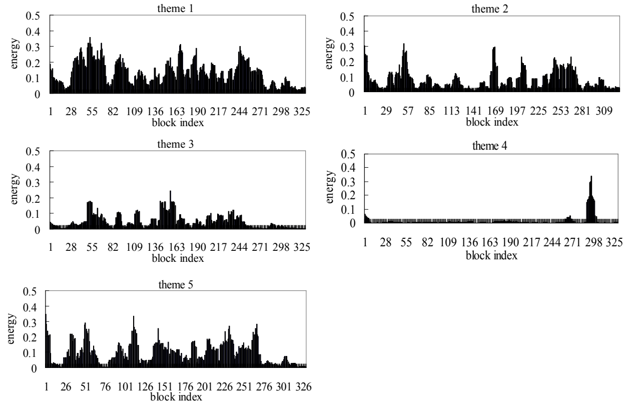

Figure 7. The energy contours of the themes in Topic 40004.

Once again, the generated evolution graph and summaries explain the storylines of the topic clearly. According to the graph, the topic was initiated by a joint exercise of submarine rescue held in the South China Sea (e_7). That exercise received a lot of attention because of the Russian submarine disaster and the fact that the Russian government was planning to launch a rescue operation a few days later. Then in events e_1 and e_8, divers began cutting holes in the Kursk to recover bodies and start the salvage process. In the meantime, the Russian government tried to determine the cause of the explosion and sinking (e_{15}, and e_{16}), and insisted that all sailors were killed instantly when the submarine sank (e_{12}). In event e_2, the divers first found three bodies of the 118 sailors. Immediately afterwards, in e_{18} and e_3, one of the bodies was identified as the sailor Kolesnikov who left a note indicating that at least 23 sailors survived for several hours after the explosion (e_4, e_{13}, and e_{18}). The note led to a great deal of public criticism about the government's slow reaction to the disaster and subsequent rescue operation. Even though officials still insisted that the sinking was caused by a collision with another submarine (e_{17}), the note forced the officials to investigate other possibilities (e_{19}). Meanwhile, the divers continued cutting holes in different compartments of the wreck to search for bodies (e_5, e_6, e_9, e_{10}, and e_{11}). During the search, Russian officials re-affirmed their determination to find the cause of the sinking and held a memorial service (e_{20} and e_{21}), even though they still speculated that it was caused by a collision with a NATO submarine (e_{22}).

As shown in Figure 7, the energy contours of Themes 1, 2, and 5 are distributed cyclically, as the events of the themes "search for bodies", "submarine salvage", and "sinking cause investigation" were mentioned frequently throughout the topic. The energy peaks also indicate the important events of the topic. For example, the peak of e_2 in Theme 1 reports the retrieval of the first bodies. The discovery of the note from the dying sailor is caught by the peak of e_{13} in Theme 3. In addition, the peaks of e_{21} and e_{14} correspond to the holding of the sailors' funerals and the drafting of the legislation, respectively. The links between events again describe the storylines of the topic well. For example, the links of Theme 5 highlight the phases of the investigation into the disaster. They sequence the theme by first mentioning that there could be a number causes of the sinking, e.g., a collision with a foreign submarine or ship, or a collision with a World War II mine. Then, the subsequent events indicate that a collision was the major cause. It is interesting to note that the contents of events e_{18} and e_{21} of the theme are relevant to the note discovery, rather than the investigation of the disaster. However, as the discovery of the note also raised hopes about learning the cause of the sinking, the investigation-related documents at that point mentioned the note many times. Thus, the theme generation process treated the note and the investigation as a single theme embedded in the inter-block association. The proposed *TS* function also extracts meaningful dependencies between events of different themes. For example, the set of edges {2, 3, 4, 5, 6, 7} effectively illustrates the associations between body recovery, submarine salvage, and the investigation. Again, the *TS* function does not allow events separated by large temporal distances to be linked, so the resulting graph is concise and comprehensible. The case studies show that the evolution graph and summaries can help users understand the storylines of topics quickly.

5. CONCLUSIONS

Topic anatomy has become an increasingly important application because of the need to grasp the gist of information contained in a large number of topic documents. Most existing summarization works try to increase the diversity of the summary to cover all the important information in the summarized documents. However, when the documents to be summarized are related to an evolutionary 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 the topic's evolution graph. Experiments based on the TDT4 corpus show that TSCAN can produce highly representative summaries that correspond well to reference summaries composed by experts. In addition, case studies show that the constructed evolution graphs accurately depict the storylines of the topics to help readers comprehend the topic quickly.

We assume that the documents are published in real-time and there is no inconsistency among the documents. The assumptions substantially reduce the difficulty of the topic evolution graph construction and summarization processes. Although the assumptions generally hold for news documents because they are written in a serious and accurate manner, it may be difficult to apply the proposed method to other unconstrained texts, such as blogs. Removing the constraints is an interesting and challenging research issue that merits further investigation.

6. ACKNOWLEDGMENTS

This work was supported in part by NSC 96-2218-E-002-040 and NSC 96-3113-H-001-012.

7. REFERENCES

[1] http://www.nist.gov/speech/tests/tdt/index.htm

[2] http://www-nlpir.nist.gov/projects/duc/index.html

[3] Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. 1998. Topic detection and tracking pilot study: final report. In DARPA broadcast news transcription and understanding workshop proceedings, 194-218.

[4] Baeza-Yates, R. and Ribeiro-Neto, B. 1999. Modern information retrieval. Addison Wesley.

[5] Chen, C. C., Chen, M. C., and Chen, M. S. 2005 LIPED: HMM-based Life Profiles for Adaptive Event Detection. In SIGKDD05 conference proceedings, 556-561.

[6] Erkan, G. and Radev, D.R. 2004. LexRank: graph-based centrality as salience in text summarization. In journal of artificial intelligence research, 22:457-479.

[7] Gong, Y. and Liu, X. 2001. Generic text summarization using relevance measure and latent semantic analysis. In SIGIR01 conference proceedings, 19-25.

[8] Kleinberg, J. 1999. Authoritative sources in a hyperlinked environment. In Journal of the ACM, 46(5):604-632.

[9] Kleinberg, J. 2002. Bursty and hierarchical structure in streams. In SIGKDD02 conference proceedings, 91-101.

[10] Lin, C.Y. and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In human language technology 2003 conference proceedings, volume 1, 71-78.

[11] Mei, Q. and Zhai, C.X. 2005. Discovering evolutionary theme patterns from text – an exploration of temporal text mining. In SIGKDD05 conference proceedings, 198-207.

[12] Nallapati, R., Feng, A., Peng, F., and Allan, J. 2004. Event threading within news topics. In CIKM04 conference proceedings, 446-453.

[13] Nenkova, A. 2005. Automatic text summarization of newswire: lessons learned from the document understanding conference. In AAAI05 conference proceedings, 1436-1441.

[14] Nomoto, T. and Matsumoto, Y. 2001. A new approach to unsupervised text summarization. In SIGIR01 conference proceedings, 26-34.

[15] Rabiner, L.R. and Sambur, M.R. 1975. An algorithm for determining the endpoints for isolated utterances. In technical journal of Bell system, 54(2):297-315.

[16] Spence, L.E., Insel, A.J., and Friedberg, S.H. 2000. Elementary linear algebra, a matrix approach. Prentice Hall.

[17] Winston, W.L. 2004. Operations research. Thomson.

[18] Yang, C.C. and Shi, X. 2006. Discovering event evolution graphs from newswires. In WWW06 conference proceedings, 945-946.

[19] Yang, Y., Pierce, T., and Carbonell, J. 1998. A study on retrospective and on-line event detection. In SIGIR98 conference proceedings, 28-36.

[20] Zha, H. 2002. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In SIGIR02 conference proceedings, 113-120.