This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

TSCAN: A Content Anatomy Approach to Temporal Topic Summarization

Chien Chin Chen and Meng Chang Chen

Abstract—A topic is defined as a seminal event or activity along with all directly related events and activities. It is represented by a chronological sequence of documents published by different authors on the Internet. In this study, we define a task called *topic anatomy*, which summarizes and associates the core parts of a topic temporally so that readers can understand the content easily. The proposed topic anatomy model, called *TSCAN*, derives the major themes of a topic from the eigenvectors of a temporal block association matrix. 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 an evolution graph of the topic. Experiments based on the official TDT4 corpus demonstrate that the generated temporal summaries present the storylines of topics in a comprehensible form. Moreover, in terms of content coverage, coherence, and consistency, the summaries are superior to those derived by existing summarization methods based on human-composed reference summaries.

Index Terms—H.2.8 [Database Applications]: Text mining, I.2.7 [Natural Language Processing]: Language summarization, I.2.7 [Natural Language Processing]: Text analysis.

1 Introduction

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 overwhelming number of documents. The situation is even more confusing if the desired knowledge is related to a temporal incident about which many independent authors have published documents based on various perspectives that, considered together, detail the development of the incident. To promote research on detecting and tracking incidents from Internet documents, the Defense Advanced Research Projects Agency (DARPA) initiated the Topic Detection and Tracking (TDT) project [1]. The project defines a topic as "*a seminal event or activity, along with all directly related events and activities*." Its goal is to detect topics automatically and track related documents from several document streams, such as on-line news feeds. The TDT project has generated a great deal of interest due to the importance and practical implications of the problem. For instance, the Google News service (see Figure 1) employs TDT techniques to organize documents related to news topics from online news websites [2][3]. While an effective TDT system can detect topics and track all related documents [1][4][5], users cannot fully comprehend a topic unless they read many of the tracked documents. For popular topics, such as "Kobe vs. LeBron in NBA MVP race" shown in Figure 1, the number of tracked documents is simply too large for users to comprehend. 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 relationships between the core parts. Applied together, the two techniques, called *topic anatomy*, can summarize essential information about a topic in a structured manner.

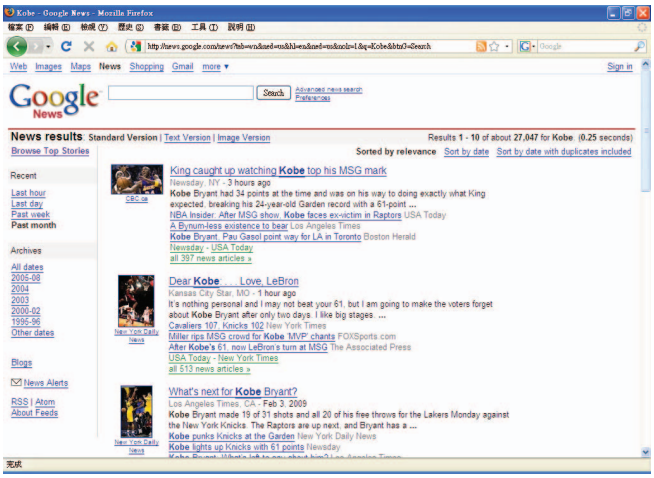

Fig. 1. Google News service (http://news.google.com)

Topic anatomy is an emerging text mining research paradigm 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 [6]. 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

- *Chien Chin Chen is with the Information Management Department, National Taiwan University, E-mail: paton@im.ntu.edu.tw (Corresponding author)*
- *Meng Chang Chen is with the Institute of Information Science, Academia Sinica, Taiwan, E-mail: mcc@iis.sinica.edu.tw*

Manuscript received (insert date of submission if desired). Please note that all acknowledgments should be placed at the end of the paper, before the bibliography.

xxxx-xxxx/0x/$xx.00 © 200x IEEE

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

topic's development [6]. 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. Events may be associated semantically 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.

Figure 2 shows a topic anatomy for the TDT4 [7] topic "President Bush Bans Abortion Funding." From the topic documents, experts identified three themes and five events. In the figure, each event is represented by a rectangle, and the events in each row represent a theme of the topic. According to the expert-generated event summaries, the three themes are "Bush's attitude to abortion", "abortion foes' opinions", and "commentators' reviews of Bush's decisions". The example demonstrates how the explicit representation of topic evolution graphs and temporal summaries can help users comprehend a topic's content and related developments easily.

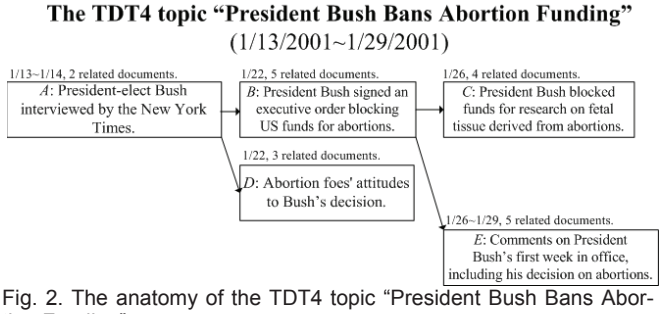

Fig. 2. The anatomy of the TDT4 topic "President Bush Bans Abortion Funding"

Although text segmentation and summarization are classic research methods in the field of information retrieval, their techniques cannot be applied to the topic anatomy problem directly without considering the temporal properties of topics. Since the content of a topic evolves along a time line, a comprehensive topic summary should highlight the topic's content and also detail the story development of the topic. In this paper, we present an anatomy-based summarizaiton method called *TSCAN* (Topic Summarization and Content ANatomy), which organizes and summarizes the content of a temporal topic described by a set of documents. TSCAN models the documents as a symmetric block association matrix, in which each block is a portion of a document, and treats each eigenvector [8] 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 then used to construct the evolution graph of the topic. The results of experiments on the official TDT4 corpus demonstrate that our anatomy-based summaries are highly representative. Moreover, they are more consistent with human composed summaries than those derived by other text summarization methods.

2 Related Work

2.1 Text Segmentation

The objective of text segmentation is to partition an input text into non-overlapping segments such that each segment is a subject-coherent unit, and any two adjacent units represent different subjects [9]. Depending on the type of input text, segmentation can be classified as *story boundary detection* or *document subtopic identification*. The input for story boundary detection is usually a text stream, e.g., automatic speech recognition transcripts from on-line newswires, which do not contain distinct boundaries between documents. Generally, naive approaches, such as using cue phrases, can identify the boundaries between documents efficiently [10]. For document subtopic identification, the input is a single document, and the task involves identifying paragraphs in the document that relate to a certain subtopic. Document subtopic identification enables many information systems to provide fine-grained services. For example, search engines can retrieve documents and return the most relevant passages segmented from the searched documents to users. Generally, the cue phrase approach is not feasible for document subtopic identification because the subtopics in a document are often similar; hence, salient cue phrases about subtopic boundaries are virtually non-existent [10][11]. Instead, decomposing a document into blocks (e.g., a set of consecutive sentences) and analyzing the word usage in each block has become a common means of identifying subtopic boundaries. However, one major problem with this approach is that the information in the blocks is usually insufficient to determine the blocks' interrelationships. Brants et al. [11] and Choi et al. [12] applied the concept of latent semantics to enrich the information in a block. Their methods use a training corpus to construct a domain-dependent thesaurus, after which the blocks are extended with their synonyms to produce better segmentation results. Blei and Moreno [13] utilized Hidden Markov Models (HMM) to detect the subtopic boundaries of a document. Under this method, the subtopics of a document are modeled as states in a HMM, and the document is treated as a series of words (or blocks), which are used to calculate the model's best state transition. A boundary occurs when two successive states in the best state transition sequence are different. Like latent semantics-based approaches, this method also requires a corpus to train the HMM's parameters, which are domain-dependent. Subsequently, Ji and Zha [10] proposed a domain-independent segmentation method that models the block associations of a document in a square matrix and treats the matrix as a gray-scale image. Then, some image processing methods are applied to sharpen the boundaries in the image. Finally, the significant and diagonal segments are selected as the paragraphs of the document.

Topic segmentation differs from document subtopic identification in a number of respects. First, the input for topic segmentation is a set of documents related to a topic, rather than a single document used in document subtopic

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

identification. Second, the identified segments of a topic, i.e., the events of themes, have a temporal property rather than a textual paragraph or several contiguous paragraphs in a document. Finally, the segments of a document are disjoint textual units, but the events of a topic can overlap temporally. In other words, several themes may be developed simultaneously. The temporal property of topic segments makes the segmentation task a challenging research issue.

2.2 Text Summarization

Generic text summarization automatically creates a condensed version of one or more documents that captures the gist of the documents. As a document's content may contain many themes, generic summarization methods concentrate on extending the summary's diversity to provide wider coverage of the content [14]. In this study, we focus on extraction-based generic text summarization, which composes summaries by extracting informative sentences from the original documents. Extraction-based text summarization methods are classified as either supervised or unsupervised. Supervised methods treat summarization as a binary-classification problem whereby each sentence of a document is labeled as "informative" or "non-informative". Shen et al. [15] proposed a supervised summarization method that employs conditional random fields (CRFs) to train a classification model, which calculates the informativeness of a sentence. Top-ranked sentences then are extracted as a summary. With appropriate training corpora, supervised summarization methods outperform unsupervised summarization methods [15]. In addition, models with large training corpora perform better than those with small training corpora [15]. However, domain-dependency is a drawback of supervised summarization methods; that is, the trained summarization model is specific to a certain document domain (e.g., sport). Deploying a supervised summarization method in a new domain involves annotating another manual training corpus, but that is a labor-intensive and time-consuming task [16][17]. In addition, number of studies [16][17][18] have shown that inter-agreement between annotators is generally low, which impacts the quality of the training corpora and the acquired summarization model. Therefore, most recently proposed summarization methods are unsupervised. Next, we consider such methods and discuss the limitations of applying them to the topic anatomy task.

Gong and Liu [14] applied Singular Value Decomposition (SVD) [8] to a document's term-sentence association matrix to perform extraction-based generic summarization. Their approach regards the decomposed singular vectors as the themes of the document and composes diverse summaries by selecting informative sentences from important themes. Nomoto and Matsumoto [16] 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. In each cluster, the sentence that contains the most information is selected as the summary. By contrast, Allan et al.'s temporal summarization method [19] processes topic sentences one by one in chronological order. The method weights the informativeness of a sentence according to the sentence's usefulness and novelty. A sentence is deemed useful for readers to comprend a topic if its content is similar to the main themes of the topic. To avoid extracting redundent summary sentences, the content of the sentence should also be different to all previously extracted sentences. In recent years, graph-based summarization methods have generated a great deal of interest [17][20][21]. Zha [17] modeled the relationships between the sentences and terms in a document as a bipartite graph. The model considers a sentence informative if it connects with many informative terms, and vice versa; and a reinforcement procedure updates the informative scores of the terms and sentences iteratively. Finally, summaries are composed by selecting informative sentences. Erkan and Radev [20] represented a set of documents as a graph in which the nodes represent sentences and the edges connect content-similar sentences. A sentence is deemed informative if it connects with several 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. Mihalcea and Tarau's method [21] constructs a graph for a document by linking similar sentence pairs in the text. Then, link analysis algorithms, such as, HITS [22] and PageRank [23], are applied to extract authoritative sentences as summaries. The approach also designates the orientation of edges in terms of the order of the sentences. Experiments show that, for news articles, graphs with backward orientations generally distill high quality summaries. This is because backward orientations tend to confer authority on early sentences, and the most important facts about a topic are usually mentioned at the beginning of a news article. To summarize a Web document, Sun et al. [24] proposed a method that collects the query terms submitted to a search engine by Internet users to retrieve documents. As contemporary search engines use link information to rank Web documents, the retrieved documents may not contain the query terms. In this case, the authors regard the query terms as implicit document content to summarize the document. The evaluation results show that query terms are representative of retrieved documents and are therefore informative sources for single document summarization. Nenkova et al. [25] conducted a series of summarization experiments to demonstrate the efficacy of using term frequencies. They showed that even a simple term-frequency approach can achieve a performance comparable to that of state-of-the-art summarization methods.

Topic summarization differs from traditional text summarization because of its temporal properties. As topics are reported chronologically, comprehensive topic summaries should describe the storylines of the topics and include informative sentences.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

2.3 Topic Evolution Mining

Kleinberg [26] developed a topic evolution mining technique that constructs a hierarchical tree from a series of topic documents. The technique utilizes a HMM-based, two-state transition diagram to model the status of topics and splits a topic into diverse themes, modeled as tree branches, if the topic contains bursty information. Nallapati et al. [27] 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 constructing a graph, Mei and Zhai [6] 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 follow the evolution of the topic and its subsequent decline. Yang and Shi [28] 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. Feng and Allan [29] proposed an incident threading method that is similar to the proposed TSCAN system. The method first identifies incidents (i.e., events) from news documents; then, the semantic dependencies between the incidents are examined to produce an incident network. The authors also defined hand-crafted rules and an optimization procedure to assign types to network links. Experiments show that link type assignment is a challenging task, and better modeling of natural languages is required to improve the technique's accuracy.

Swan and Allan [30] proposed a timeline system to display important topics in a document corpus graphically. The system utilizes a statistical feature selection method to identify terms that occur frequently in a specific time period. The identified term-period groups, i.e., topics, are then organized sequentially to form the timeline of the corpus. This timeline system can be applied to a document corpus, so the result is similar to that of the TDT project. In contrast, we focus on a single topic and documents specifically related to that topic.

3 TSCAN System

In this section, we present our model and the methods used in the proposed topic anatomy system.

3.1 Topic Model

A *topic* is a real world incident that comprises one or more *themes*, which are related to a finer incident, a description, or a dialogue about 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. Naturally, all the events taken together form the storyline(s) of the topic. Although the events of a theme are temporally disjoint, they are considered semantically dependent in order to express the development of the theme. Moreover, events in different themes may be associated because of their temporal proximity and context similarity. The proposed method identifies themes and events from the topic's documents, and connects associated events to form the topic's evolution graph. In addition, the identified events are summarized to help readers better comprehend the storyline(s) of the topic. Figure 3 illustrates the relationships between the themes, events, and event dependencies of a topic in the proposed model.

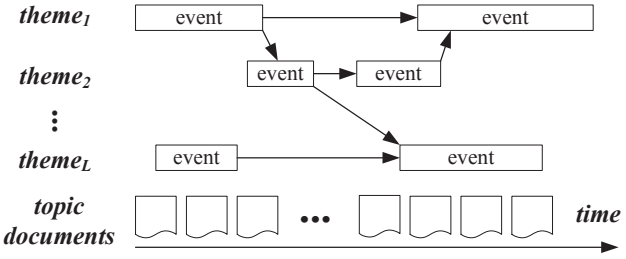

Fig. 3. The relationships between themes, events, and event dependencies

A topic is represented explicitly by a collection of chronologically ordered documents. In this study, we assume that the documents are published in the same order as the events of the topic reported by independent authors, and that there is no inconsistency between the contents of the documents. TSCAN decomposes each document 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. For a topic, let $T = \{t_1, t_2, \ldots, t_m\}$ be a set of stemmed vocabulary without stopwords [31]. The topic can then be described by an m x n *term-block association matrix* B in which the columns $\{b_1, b_2, \ldots, b_n\}$ represent the blocks decomposed chronologically from the topic documents. In other words, for any two blocks, b_i and b_j, if $i < j$, then either the document containing b_i was published before the document containing b_j, or b_i appears before 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 [31].

3.2 Theme Generation

A 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 in 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 the constitution of a vector $\underline{v}$, $\underline{v}^TA\underline{v}$ computes the theme's association to the topic's content. The objective function 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.

$$\max \underline{v}^TA\underline{v} \quad (1)$$

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

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 to the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

set of normalized vectors such that the following Lagrangian formula [32] can be used to solve Eqs. (1) and (2).

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

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

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

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

Equation (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 n x n square matrix, there are at most n eigenvectors [8]. In terms of non-linear programming, Eq. (3) can have more than one stationary point [32]. To derive the relevant themes of the topic, we employ the following theorem for symmetric matrices.

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 $V = \{\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_n\}$ 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 discussion of diagonalization of symmetric matrices in many linear algebra books [8]. Since V is an orthonormal basis of R^n, its inverse is identical to its transposition, i.e., $V^{-1} = V^T$ [8]. Therefore, the matrix A can be represented as follows:

$$\begin{aligned} A &= VDV^{-1} = VDV^T \\ &= [\underline{v}_1, \ldots, \underline{v}_n][d_{1,1}\underline{e}_1, \ldots, d_{r,r}\underline{e}_r, 0\underline{e}_{r+1}, \ldots, 0\underline{e}_n]V^T \\ &= [d_{1,1}\underline{v}_1, \ldots, d_{r,r}\underline{v}_r, 0\underline{v}_{r+1}, \ldots, 0\underline{v}_n][\underline{v}_1, \ldots, \underline{v}_n]^T \\ &= d_{1,1}\underline{v}_1\underline{v}_1^T + \ldots + d_{r,r}\underline{v}_r\underline{v}_r^T + 0\underline{v}_{r+1}\underline{v}_{r+1}^T + \ldots + 0\underline{v}_n\underline{v}_n^T, \end{aligned} \quad (6)$$

where $\underline{e}_i$ denotes the standard vectors of R^n [8]. In other words, the symmetric matrix A can be decomposed into the sum of n matrices spanned by its eigenvectors. We treat the first significant L ($L<r$) eigenvectors of A as the themes of the topic. Then, the inter-block association approximated by the selected themes can be represented as follows:

$$\begin{aligned} A &\approx d_{1,1}\underline{v}_1\underline{v}_1^T + d_{2,2}\underline{v}_2\underline{v}_2^T + \ldots + d_{L,L}\underline{v}_L\underline{v}_L^T \\ &= [\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_L][d_{1,1}\underline{e}_1, \ldots, d_{L,L}\underline{e}_L][\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_L]^T \\ &= V_L D_L V_L^T, \end{aligned} \quad (7)$$

where V_L, called *theme matrix*, is an n x L matrix in which a column represents a theme; and D_L is an L x L diagonal matrix whose 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 derived themes tend to be unique and descriptive.

Actually, the eigenvectors used in V_L are identical to the top L right singular vectors of B, ranked by their singular values. However, as the dimension of A is generally much smaller than that of B, i.e., $n << m$, calculating V_L by using Theorem 1 for symmetric matrices is faster and requires less memory than using SVD. Moreover, in contrast to probability-based counterparts of SVD, e.g., PLSA [33], where different numbers of themes may result in a distinct collection of themes, the eigenvectors of A are always the same and are irrelevant to the value of L. Therefore, in our approach, L is a free parameter that allows users to decide a suitable granularity for topic anatomy.

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 taken as a noteworthy event embedded in the theme, and valleys (i.e., a sequence of small values) in $\underline{v}_j$ may be event boundaries. However, according to the definition of eigenvectors, the signs of entries in an eigenvector are invertible. Moreover, Kleinberg [22] and Nicholas and Dahlberg [34] showed that both the positive and negative entries of an eigenvector contain meaningful semantics for describing a certain concept embedded in a document corpus; and the amplitude of an entry determines the degree of its correlation to the concept. Note that the tasks of our event segmentation and speech endpoint detection are similar in that they both try to identify important segments of sequential data. In addition, it is the amplitude of sequential data that determines the data's importance. For example, given the speech utterance in Figure 4, the speech endpoint detection task involves distinghishing the significant segment S_2 from the insignificant silent segments (S_1 and S_3) mixed with background noise. Here, S_2 represents the word 'one' and comprises a sequence of points with large positive and negative amplitudes. Therefore, we adopt Rabiner and Sambur's R-S endpoint detection algorithm [35] for event segmentation. To segment events, the R-S algorithm examines the amplitude variation of an eigenvector to find the endpoints that partition the theme into a set of significant events.

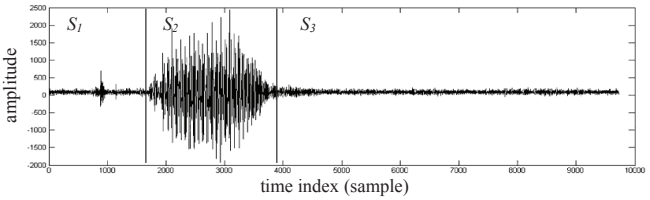

Fig. 4. The waveform of an utterance of the word 'one'

In the R-S algorithm, every block in an eigenvector has an energy value. To calculate the energy, we adopt the square sum scheme, which has proved effective in detecting endpoints in noisy speech environments [36] and is defined as follows:

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

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

Figure 5 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 se-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

quence 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 thd_{seg} as 0.1; then, we scan the energy contours linearly to find consecutive blocks whose energy values are above the threshold. To reduce the number of false alarms during event segmentation and refine the segmentation result, we employ the two frequently used heuristics: 1) we merge two close events, and 2) we prune small events [36]. 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 because of 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 and evolution of a theme.

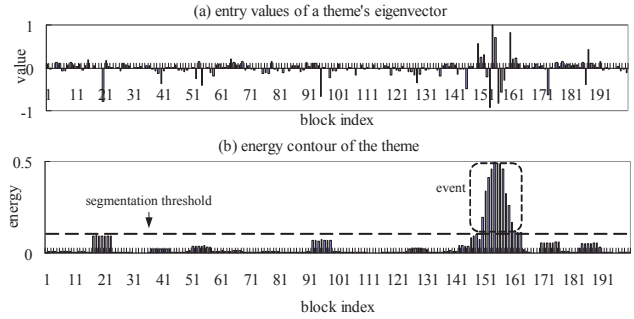

Fig. 5. The eigenvector of a theme and its energy contour

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 [14][16][17][19][20][25] try to cover diverse themes in document summaries; however, our method further describes the development of themes via summarized events to help users comprehend a topic's storylines.

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 a 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 let $<e_k.fb, e_k.lb>$ be the event's timestamp, where $e_k.fb$ and $e_k.lb$ are the indexes of the first and last blocks, respectively; and $|e_k| = 1 + e_k.lb - e_k.fb$ 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 $e_j.fb > e_i.fb$.

Automatic induction of event dependencies is often difficult due to the lack of sufficient domain knowledge and effective knowledge induction mechanisms [27][29]. However, as event dependencies usually involve similar contextual information, such as the same locations and person names, they can be identified through word usage analysis [6][27][28][29]. Our approach, which is based on this rationale, involves two procedures. First, we link events segmented from the same theme sequentially to reflect the theme's development. Then, we use a temporal similarity function to capture the dependencies of events in different themes. For two events, e_i and e_j, belonging to different themes, where $e_j.fb > e_i.fb$, we calculate their *temporal similarity* (*TS*) by Eq. (9):

$$TS(e_i,e_j) = TW(e_i,e_j) * cosine(e_i.cv,e_j.cv), \quad (9)$$

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

$$e_k.cv = \frac{1}{|e_k|}\sum_{i=e_k.bb}^{ek.eb} |v_{i,e_k.ev}| * \underline{b}_i, \quad (10)$$

where $\underline{b}_i$ is the term vector of block i. In short, $e_k.cv$ averages the term vectors of the event's blocks in accordance with their correlation to the event. The *temporal weight* (*TW*) function, defined in Eq. (11), then weights the 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.

$$TW(e_i,e_j) = \begin{cases} 1-\frac{e_j.bb-e_i.eb}{n}, & \text{if } e_j.bb>e_i.eb, \\ 1-\frac{2*(\min(e_i.eb,e_j.eb)-e_j.bb)}{|e_i|+|e_j|}, & \text{if } e_j.bb\le e_i.eb. \end{cases} \quad (11)$$

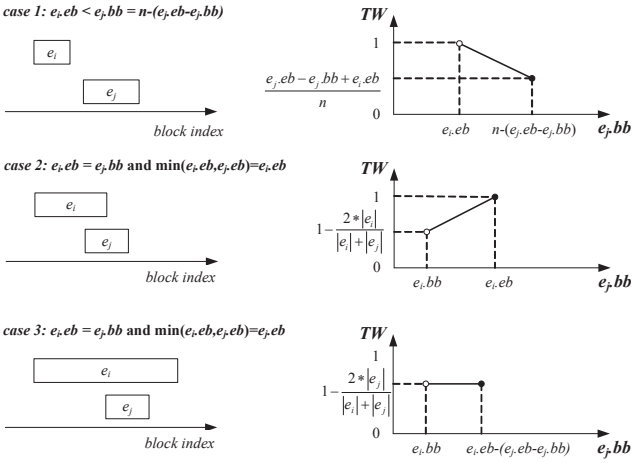

Fig. 6. The graphs of the *TW* function under three temporal cases

The range of the proposed *TW* function is within (0, 1]. As shown in Figure 6, *TW* considers the temporal relationship between events e_i and e_j and assigns an appropriate temporal weight. In case 1, where e_i and e_j do not overlap, *TW* penalizes the events with a large temporal distance. The penalty corresponds to Yang's observation [4] that temporally close pieces of information are usually more relevant to one another than those farther apart. In case 2, e_i and e_j do overlap and e_j is not contained in e_i. *TW* also penalizes events if their initial timestamps are close together. The penalty is based on the supposition that when two events occur almost simultaneously, they are probably distractions caused by 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 $|e_j|$. As in case 2, the property pre-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

vents the linking of events with similar timestamps because they may be distractions caused by a prior event.

4 Performance Evaluations

4.1 Data Corpus

In [37], two case studies using the official TDT topics are provided to demonstrate that the evolution graphs constructed by TSCAN can extract the themes, events, and event dependencies of the examined topics successfully. In this research, we evaluate our anatomy-based summarization technique by comparing the derived summaries with those of several text summarization methods. We use the official TDT4 topics for the performance evaluations. The Linguistic Data Consortium has compiled a series of TDT corpora for the annual TDT contests. The TDT4 corpus comprises 28,390 English news documents from eight well-known news agencies for the period October 1, 2000 to January 31, 2001. Among them, seventy news events with 1,926 related documents were labeled by NIST annotators for various TDT evaluation tasks. The annotators also composed factual descriptions of the topics, which are regarded as human-composed reference summaries for summarization evaluations. Although Document Understanding Conferences[1] (DUC) also use TDT topics for summarization contests, the average size of the topics is only 10 documents, which is too small for the purpose of topic anatomy. We therefore select 26 TDT4 topics, each containing more than 20 documents, for evaluation. Table 1 details the evaluated topics in our data corpus.

TABLE 1
STATISTICS OF EVALUATED TOPICS

Number of topics	26
Number of news documents	1,211
Average number of documents per topic	46.6
Number of sentences	32,739
Average number of sentences per topic	1,259.2

In the pre-processing phase, each topic document is partitioned into blocks of sentences by using a simple Perl script[2] supplied by DUC. The the system parameters H and w represent, respectively, the length of the sliding window used to aggregate the energy of a block in the R-S endpoint detection algorithm and the number of sentences in a block. To assess the influence of H and w on the summarization performance, they are set at {5, 7, 9} and {1, 3, 5} respectively. The parameter L is critical to the quality of detected themes. The function $U(L)$, defined in Eq. (12), is used to measure the underestimation of V_L when approximating the inter-relations of a topic's blocks.

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

$U(L)$ is the average of the squared differences between A and $V_L D_L V_L^T$. A low $U(L)$ value indicates that the selected L themes represent the inter-block associations sufficiently well. From Eq. (6), it is clear that the larger the number of themes selected, the lower will be the value of $U(L)$. However, a large L may be a drawback because the constructed evolution graph may have too many themes to be comprehensible. For summarization comparison, the evaluations are performed with L = 1 to 10 in order to illustrate the influence of themes on the summarization performance. Figure 7 shows the average $U(L)$ of the 26 evaluated topics for L = 1 to 10. It is noteworthy that, contrary to expectations, blocks with little content information (i.e., w = 1) produce a low underestimate. This is because the block association matrix constructed with small-size blocks is very sparse. Thus, the slight difference between A and $V_L D_L V_L^T$ reduces the value of $U(L)$ substantially.

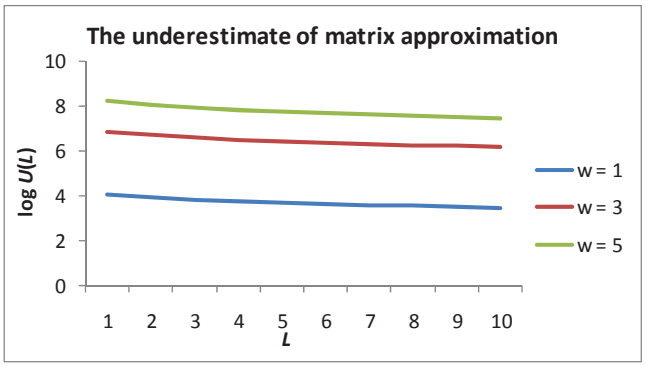

Fig. 7. Underestimation of L themes

4.2 Summarization Evaluations

We compare the summarization performance of TSCAN with the following six well-known 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. 3) The SVD method [14], which composes summaries by extracting the blocks with the largest entry value in singular vectors. Note that the result derived by the SVD method is identical to that of the graph-based summarization method [20]. 4) The K-means method [16], which compiles summaries by selecting the most salient blocks of the resulting K clusters. Generally, this method's performance depends on the quality of the initial clusters. In our experiments, to ensure fair comparison of the K-means method, we use the best result from fifty randomly selected initial clusters for evaluation. 5) The temporal summary (TS) method [19], where we adopt the $useful_2$ and $novel_1$ techniques proposed by the authors to compute the informativeness score of a topic block. We do not adopt the $novel_2$ technique because the authors have shown that the performance difference between using $novel_1$ and using $novel_2$ is not significant. In addition, $novel_2$ requires a training corpus to derive an appropriate number of clusters (i.e., parameter m), but the training corpus is not availabe. 6) The frequent content word (FCW) method [25], which constructs summaries by selecting blocks with frequent terms. This method's performance is comparable to that of state-

[1] http://duc.nist.gov/
[2] http://duc.nist.gov/duc2004/software/duc2003.breakSent.tar.gz

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

of-the-art summarization methods. In addition, we adopt Nenkova et al.'s context adjustment technique to increase the summary diversity.

We use four metrics, the summary-to-document content similarity (SDCS), average pairwise summary block similarity (APSBS), ROUGE [38], and execution time, to evaluate the above summarization methods. The SDCS metric compares the content coverage of a generated summary to that of the documents selected for summarization; APSBS measures the degree of content coherence of a summary; ROUGE considers the consistency between the content of a generated summary and that of a set of expert-composed reference summaries; and execution time measures the efficiency and scalability of the summarization methods. Naturally, it is preferable that the summary should be content-coherent and as close as possible to the reference summaries. It should also have a high degree of content diversity to cover all the important information in the summarized documents.

The summarization evaluation procedure involves the following steps. First, for each setting of the parameters L, w, and H, we 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 to produce summaries of the same size (in terms of the number of blocks) as those generated by TSCAN. For instance, if our summary for an evaluated topic consists of z blocks, the summaries of the forward and backward methods will contain the first and last z topic blocks respectively. The TS method compiles a summary by extracting the first z useful and novel blocks. In the SVD method, the K-means method, and the frequent content word method, summaries are compiled by selecting the most salient block from each of the top z singular vectors, z clusters, and z term frequencies respectively. Table 2 shows number of blocks accumulated for summaries under different parameter settings and the corresponding compression ratios.

TABLE 2
THE NUMBER OF BLOCKS ACCUMULATED FOR SUMMARIES

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

Legend: (x%) represents the compression ratio of the summary and x = 1 – (summary size / topic size).

Overall, the compression ratios of the evaluated summaries are high, as at least 87% of the topics' contents are omitted. In addition, the smaller the number of themes used to construct topic summaries, the higher will be the compression ratios. Under the proposed method, the major theme of a topic contributes more events for summarization than the remaining less significant themes. This is similar to reference summary composition, where topic annotators usually highlight the events of the main theme. Table 2 also shows that parameter H plays an important role in the event segmentation process. When H = 5, the energy contours are not smooth enough and contain many small sawteeth. Thus, most segmented events have short spans and are eliminated by the R-S algorithm. Conversely, H = 9 overly smoothes the energy contours, so many adjacent events merge. Setting H at 7 yields low compression ratios because it balances the tradeoff between H = 5 and H = 9, and achieves superior performances in our summarization experiments.

4.2.1 Summary-to-document Content Similarity

Summary-to-document content similarity is defined as the average cosine similarity between an evaluated summary and the topic documents. Both components 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. Figure 8 shows the micro average summary-to-document content similarity scores derived by the compared methods. As shown in the figure, our method outperforms the compared methods with small L values. When w = 1, the TS method yields superior SDCS scores. This is because the $useful_2$ technique determines the informativeness of a block by calculating the probability that the block is generated by the language model of the block's document. Therefore, the selected summary blocks are highly similar to the documents they are extracted from, which increases the SDCS scores. It is intereting to note that the SDCS score of TS decreases as w increases. This is because topic documents generally report the latest information about a topic, so the contents of the docments would be different. Even though the summary blocks of the TS method are highly similar to the documents they are extracted from, they are somewhat dissimilar to the other topic documents. A large w increases the content of a block and this also magnifies the dissimilarity. The K-means method achieves a higher similarity score because its summary provides better coverage of the topic's contents. Our method simply selects the top L significant themes (L<<r) to represent a topic, whereas the K-means method partitions all of a 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 cluters 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; therefore, they would be difficult for users to understand [37]. Compared to the SVD method, which is also a vector-based summarization method, the superior SDCS scores achieved by our method demonstrate the advantage of using event

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

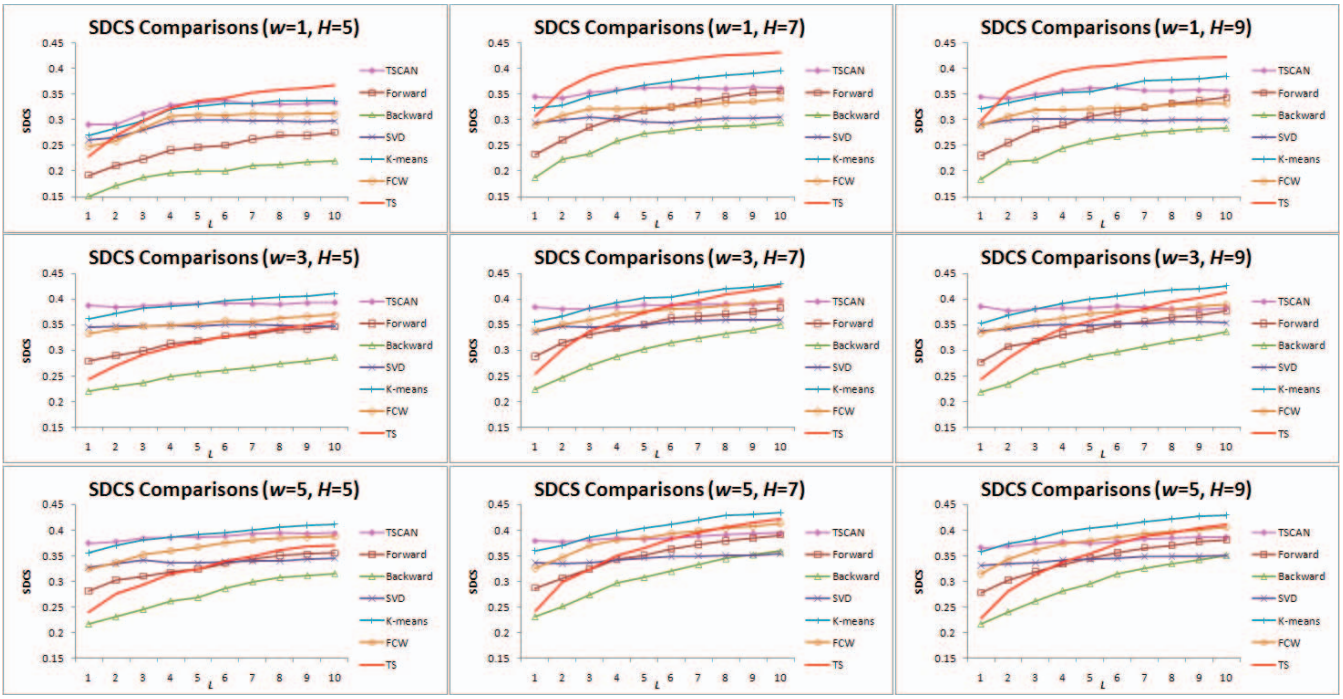

Fig. 8. Micro average summary-to-document content similarity

segmentation for temporal topic summarization. The SVD method does not consider the temporal information (i.e., events) of topics and strives to increase the diversity of summaries by including a lot of side information extracted from minor singular vectors. As a result, the constructed summaries deviate from the core content of the topics, which impacts the SDCS performance. The summaries extracted by the frequent content word method are based on a multinomial model constructed from all the topic documents. Thus, similar to the *K*-means method, the summaries cover the topic contents entirely and are even better than our summaries under a few experiment settings. Intuitively, the coverage provided by the forward and backward methods is poor because their summaries only cover the beginning and end of each topic respectively. To improve their performance, lower compression ratios are required so that more topic information can be included in the summaries. Note that for low compression ratios, the SDCS scores of the forward and backward methods are overestimated. This is because the blocks extracted by the methods cover the first few topic documents and the last few topic documents respectively. The content similarity of the summaries and the documents is thus extremely high and biases the SDCS performance.

To assess the effect of the system parameters on TSCAN's summarization performance, we average TSCAN's SDCS scores under each setting of the parameters H and w. Parameter H, which determines the degree of smoothing in event segmentation, affects the number of segmented events and summary size, as shown in Table 2. Figure 9 illustrates that H has little effect on SDCS performances. Thus, as long as the summary covers the core of a topic, small summaries will achieve similar SDCS performances to those of large summaries. Parameter w controls the granularity of topic blocks. In the pre-

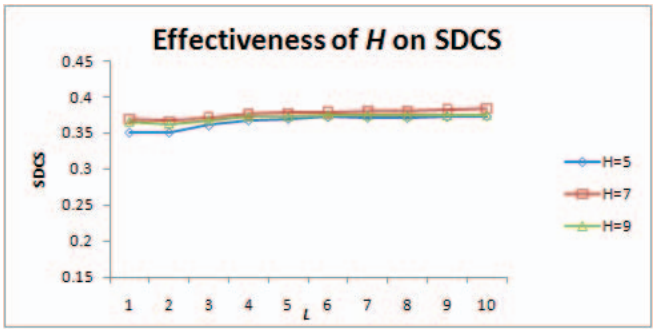

Fig. 9. The effect of parameter *H* on the SDCS metric's performance

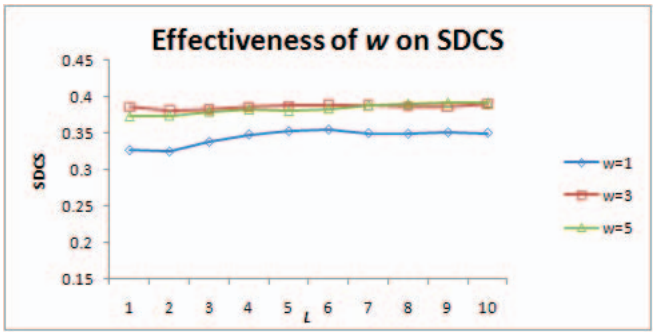

Fig. 10. The effect of parameter *w* on the SDCS metric's performance

processing phase of the experiments, we observed that the sentence segmentation program supplied by DUC sometimes segments sentences incorrectly when dealing with noun abbreviations followed by a period. As a result, the setting w = 1 yields inferior SDCS performances because each summary block contains a short segmented sentence that may not convey complete topic information. The nearly equal SDCS performance scores under w = 3 and w = 5 also show that the summary size or block size does not affect the coverage of the extracted summaries

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

provided that the extracted blocks contain enough topic information.

4.2.2 Average Pairwise Summary Block Similarity

Although our summaries are not as diverse as those of the *K*-means method, they are more coherent. A popular measurement frequently used to judge the content coherence of a set of documents is the *average pairwise document similarity* [3]. The metric models documents as term vectors and computes the similarity between documents in terms of the cosine similarity. As the cosine similarity calculates the degree of word overlap between documents, a high average pairwise similarity indicates that the documents have a significant amount of word overlap; hence, the documents are content-coherent. We employ the measurement to calculate the degree of content coherence of a summary. For an evaluated summary, which is represented by a set of summary blocks, we model each summary block as a TF-IDF term vector (i.e., $\underline{b_i}$) and compute the average cosine similarity of all pairs of block vectors. We then average the average pairwise summary block similarity (APSBS) of the compared methods under various parameter settings. The results are shown in Figure 11.

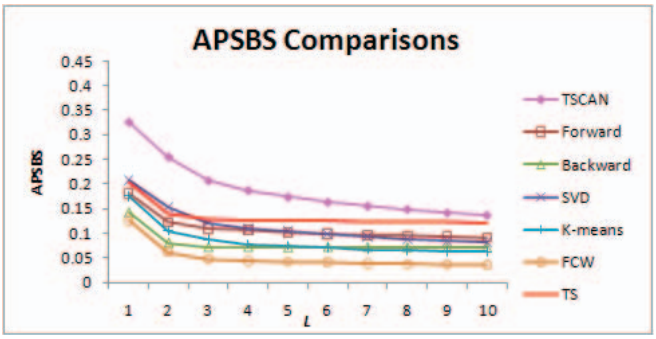

Fig. 11. Comparison of the average APSBS scores

As shown in the figure, TSCAN achieves superior APSBS scores. The reason is that our summaries focus on events in the first few (i.e., top L and L <= 10) significant themes; therefore, summary blocks have similar contexts. By contrast, the summaries compiled by other approaches try to cover diverse themes, so they are less coherent than our summaries. For all the summarization methods, APSBS decreases as the size of summaries increases (i.e., L increases). As a large summary covers many themes, its content is more diverse than that of a small summary. Hence, the average pairwise similarity will be low. FCW produces the lowest average pairwise similarity because the method's context adjustment mechanism reduces the weight of a word if it was included in the summary during the summarization process. The objective is to avoid repeating information in the summary. As a result, the selected summary blocks have little content overlap, so the APSBS score is low. It is noteworthy that, overall, the forward method produces good APSBS. The forward method selects the initial blocks of a topic as the topic summary. Generally, the first few blocks of news documents introduce the main themes of a topic in a comprehensible manner. The content of the blocks is context-coherent (i.e., each block mentions the same named entities) to help readers comprehend the topic. Therefore, the content overlap is high so the APSBS are good.

4.2.3 ROUGE Evaluations

ROUGE is a recall-oriented summary evaluation metric that is widely used in DUC contests. It measures the summarization performance by calculating the number of overlapping n-grams [3] between an evaluated summary and a set of reference summaries. Note that, in our study, the reference summaries are the topic explanations composed by NIST annotators. The ROUGE score is 1 if the evaluated summary is consistent with the reference summaries; and 0 otherwise. It has been shown that the results of comparisons based on ROUGE-1 and ROUGE-2 (i.e., unigram- and bigram-overlap) are highly consistent with those derived by human evaluators [15][38]. Therefore, we use ROUGE-1 and ROUGE-2 to evaluate the consistency of manual summaries derived by the compared methods.

Figures 12 and 13 show the micro average performances of ROUGE-1 and ROUGE-2 respectively. The proposed method achieves the best ROUGE-1 and ROUGE-2 scores under nearly all parameter settings. As ROUGE is a recall-oriented evaluation metric, the scores of all the compared methods increase as L increases. It is interesting that the improvement achieved by our method over the compared methods increases as the compression ratio increases (i.e., L decreases). For example, when L = 1, our method outperforms the compared methods by 7.2% to 85.9% for ROUGE-1 and by 7.5% to 163.1% for ROUGE-2. Moreover, for compression ratios > 95%, the improvements over the compared methods are statistically significant with a 99% confidence level in a one-tail paired t-test. The superior ROUGE performance of TSCAN is related to the unique feature of our summary composition mentioned in Sec. 3.3. The *K*-means method, SVD method, TS method, and frequent content word method focus on increasing summary coverage by using clusters, singular vectors, block novelty, and context adjustment respectively. In contrast, our method distinguishes between important events in themes to achieve both summary diversity and narrative tracing properties. The results of TSCAN are thus consistent with topic annotators' reference summaries, which usually explain the major events of significant themes and sometimes include other items of information to describe the storylines of a topic. As a result, TSCAN outperforms the compared methods in terms of the ROUGE evaluation metric. Nenkova [39] regards the backward method as effective because its ROUGE performances have proved comparable to those of many elaborate summarization methods in a number of contests. However, based on our evaluations, the method is inferior to the simple forward method. Mihalcea and Tarau [21] observe that the forward approach is effective in summarizing news documents because of the style of news report writing. The first few sentences of a news document usually describe the gist of the story. Thus, the forward method is more effective. In fact, it is nearly as good as many of the compared methods under certain parameter settings. It is interesting to note that the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

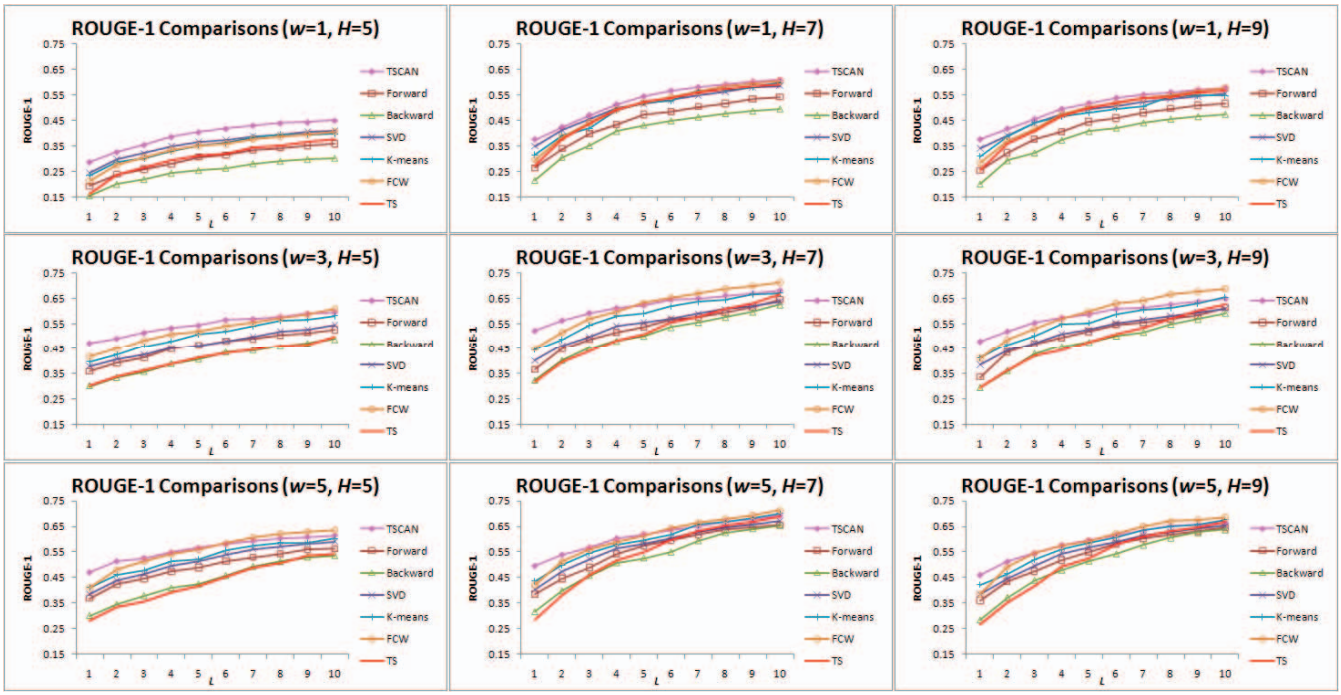

Fig. 12. ROUGE-1 comparisons

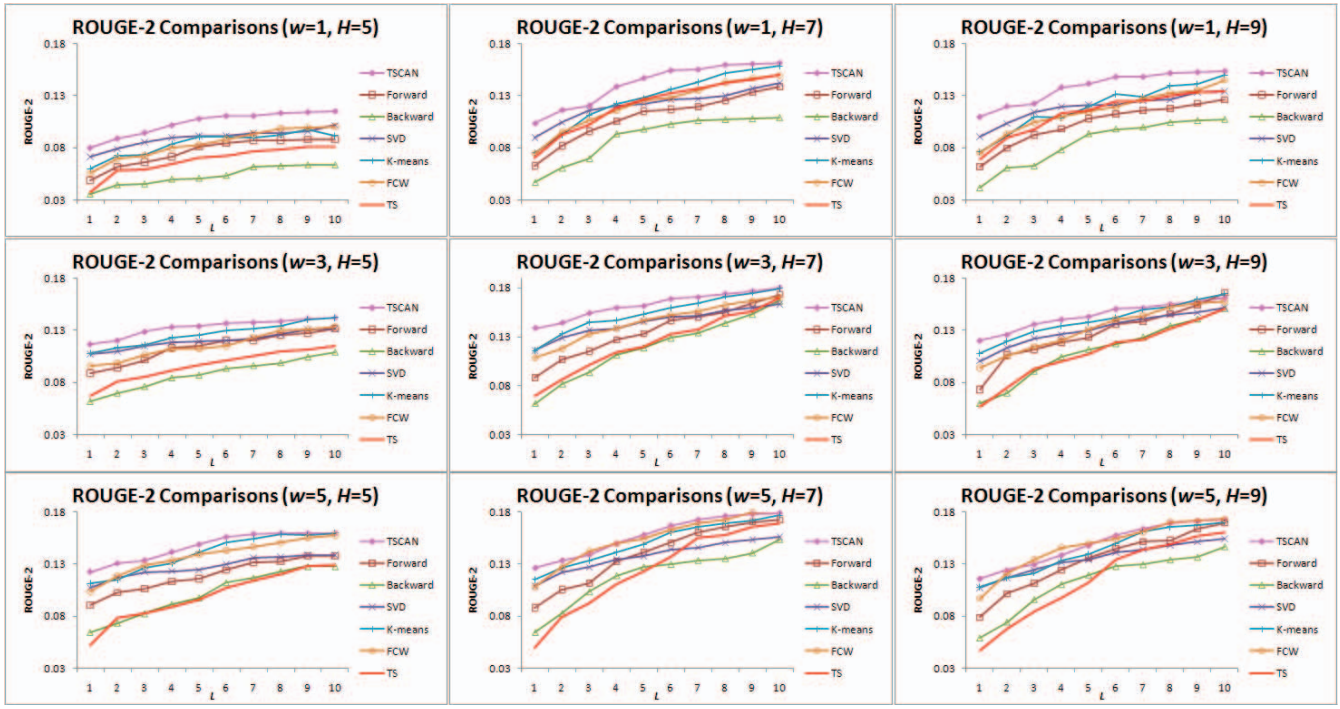

Fig. 13. ROUGE-2 comparisons

ROUGE performances of the TS method are inferior. This is because the method tries to extract the first blocks that mention new topic events. Although the extracted summary blocks possess novel information, they may not cover the core of the topic events; hence the ROUGE scores are low. The frequent content word method achieves comparable ROUGE-1 performances, but its ROUGE-2 performances are not impressive. The method is based on a multinomial model, which assumes that words are independent of each other [3]. However, since the assumption is contrary to the definition of a bigram, the ROUGE-2 performance deteriorates.

Figures 14 to 17 show the effects of the system parameters on TSCAN's ROUGE performance. Unlike the outcomes we observed under SDCS, the size of a summary affects our method's ROUGE performance. This is because ROUGE is a recall-oriented evaluation metric, so the score is generally proportional to the size of the summaries. Therefore, setting H at 7 produces good ROUGE-1 and ROUGE-2 performances as the corresponding compression ratio is low. Again, setting w at 1 generates small and incomplete summary blocks that impact ROUGE per-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

formances. While w = 5 produces better ROUGE-2 performances than w = 3 for large summaries, the overall difference in ROUGE performances under the two parameter settings is not significant.

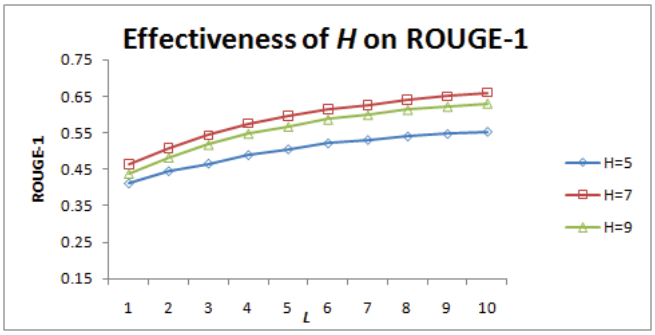

Fig. 14. The effect of parameter H on ROUGE-1 performances

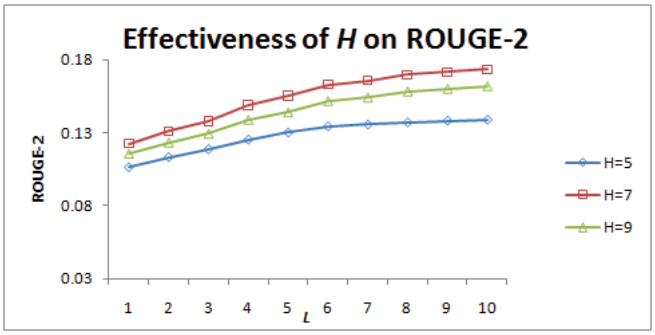

Fig. 15. The effect of parameter H on ROUGE-2 performances

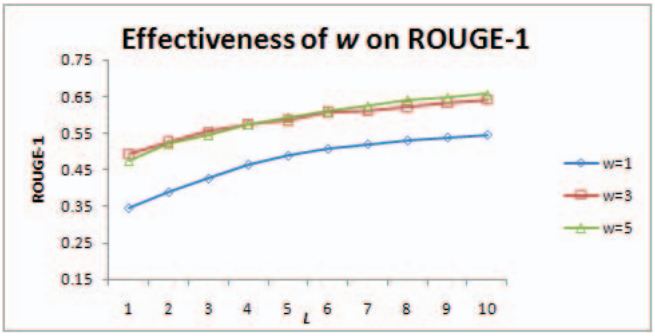

Fig. 16. The effect of parameter w on ROUGE-1 performances

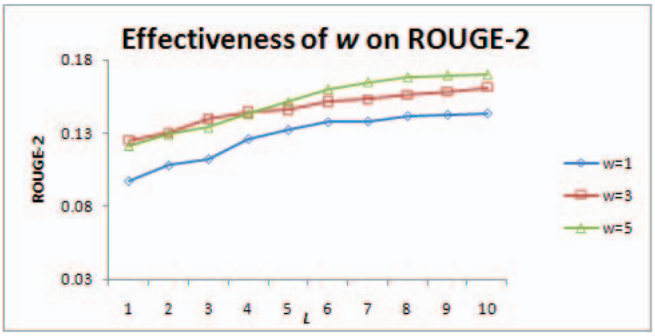

Fig. 17. The effect of parameter w on ROUGE-2 performances

4.2.4 Scalability and Time Comparisons

We evaluated the execution time of the compared summarization methods on an AMD AthlonTM 64 Processor 3200++ PC with the Windows XP Service Pack 3 operating system and a 2GB main memory. For each method, we recorded the time required to generate the summaries of the 26 evaluated topics under a specific parameter setting. However, due to space limitations, we only show the average execution times of the methods under all parameter settings in Figure 18. We do not show the execution times of the forward and backward methods because they do not need to weight each topic block to compile topic summaries. Therefore, their respective execution times are constant and irrelevant to the parameter settings. FCW is an iterative summarization method. In each iteration, it computes the weight of every block based on a topic model; therefore, its time complexity is O(n). K-means is also a linear algorithm [3]. Although the TS method computes the weight of topic blocks linearly, it needs to examine the content of previous blocks to compute a block's novelty; therefore, its time complexity is O(n^2). The SVD method and our method spend much of the computation time calculating eigenvectors. We employ MATLAB to calculate a matrix's eigenvectors. The time complexity is O(n^2I^2), where I is the number of eigenvectors to be computed [40]. The results in Figure 18 show that the linear summarization methods (i.e., K-means and FCW) run faster than the other methods. In terms of time complexity, the eigen-based methods (i.e., the SVD method and our method) run slower than the TS method. However, we observe that when w = 3 and w = 5 our method runs faster than the TS method. We believe that the longer execution time of the TS method is due to program implementation issues, as MATLAB is a commercial software package implemented by experienced programmers. Except for the TS method, the execution time of the compared methods generally increases as the size of the summary (i.e., L) increases. For the SVD method and TSCAN, a large L means that the methods need to examine a lot of eigenvectors to compile topic summaries; therefore, the execution time increases. For the K-means method and the FCW method, a large L increases the number of clusters (i.e., K) and the number of iterations required to extract summary blocks with frequent terms respectively. Hence, the methods' execution times also increase. It is noteworthy that the TS method's execution time is irrelevant to the size of the summary. This is because the method must weight all topic blocks irrespective of how many summary blocks are required to compile a topic summary.

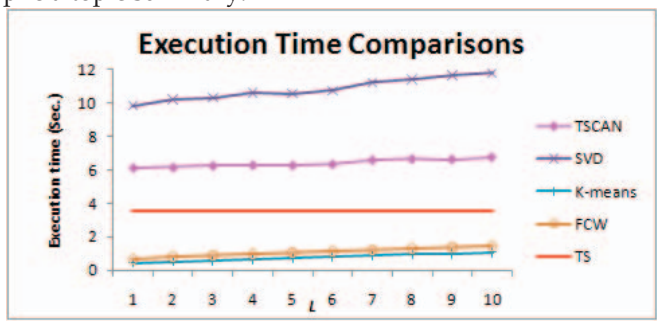

Fig. 18. Comparison of execution times

To evaluate the scalability of TSCAN, we show the time it requires to generate the summaries of the 26 topics in Figure 19. The values of parameters w, h and L are set at 1, 7, and 10 respectively, as this setting requires the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

longest execution time in our experiments. As shown in the figure, for most of the topics, TSCAN requires no more than 20 seconds to compile topic summaries. For the largest topic which consists of 128 topic documents and 4527 topic sentences, TSCAN takes approximately 2 and half minutes to construct the topic summary. In practice, the computation time might be less than the time required to crawl the topic documents. Thus, the proposed method is feasible for real-world summarization systems.

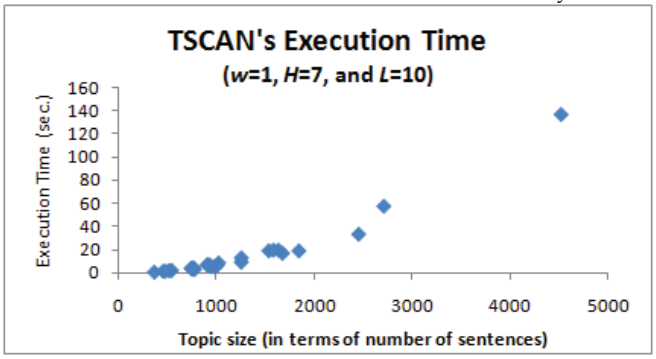

Fig. 19. The scalability of TSCAN

4.2.5 Performance Review

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

6 Conclusions

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

In this paper, we have presented a topic anatomy system called TSCAN, which extracts themes, events, and event summaries from topic documents. Moreover, the summarized events are associated by their semantic and temporal relationships, and presented graphically to form an evolution graph of the topic. Experiments based on official TDT4 topics demonstrate that TSCAN can produce highly representative summaries that correspond well to the reference summaries composed by experts.

References

[1] Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. "Topic Detection and Tracking Pilot Study: Final Report," in Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998, 194-218.

[2] Hatzivassiloglou, V., Gravano, L., and Maganti, A. "An investigation of linguistic features and clustering algorithms for topical document clustering," in Proceedings of the 23rd Annual international ACM SIGIR Conference on Research and Development in information Retrieval, 2000, 224-231.

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

[4] Yang, Y., Pierce, T., and Carbonell, J. "A Study on Retrospective and On-Line Event Detection," in Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, 1998, 28-36.

[5] Chen, C.C., Chen, M.C., and Chen, M.S. "An Adaptive Threshold Framework for Event Detection Using HMM-based Life Profiles," in ACM transactions on information systems, volume 27, issue 2 (2009).

[6] Mei, Q., and Zhai, C. X. "Discovering Evolutionary Theme Patterns from Text – An Exploration of Temporal Text Mining," in Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005, 198-207.

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

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

[9] Hearst, M. A., and Plaunt, C. "Subtopic Structuring for Full-Length Document Access," in Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, 1993, 59-68.

[10] Ji, X., and Zha, H. "Domain-independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming," in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, 2003, 322-329.

[11] Brants, T., Chen, F., and Tsochantaridis, I. "Topic-Based Docu-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

ment Segmentation with Probabilistic Latent Semantic Analysis," in Proceedings of the eleventh international conference on Information and knowledge management, 2002, 211-218.

[12] Choi, F. Y. Y., Wiemer-Hastings, P., and Moore, J. "Latent Semantic Analysis for Text Segmentation," in Proceedings of 2001 conference on Empirical Methods in Natural Language Processing, 2001, 109-117.

[13] Blei, D. M., and Moreno, P. J. "Topic Segmentation with an Aspect Hidden Markov Model," in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001, 343-348.

[14] Gong, Y., and Liu, X. "Generic Text Summarization Using relevance Measure and Latent Semantic Analysis," in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001, 19-25.

[15] Shen, D., Sun, J.T., Li, H., Yang, Q., and Chen, Z. "Document Summarization Using Conditional Random Fields," in proceedings of the 20th international joint conference on artificial intelligence (2007), 2862-2867.

[16] Nomoto, T., and Matsumoto, Y. "A New Approach to Unsupervised Text Summarization," in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001, 26-34.

[17] Zha, H. "Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering," in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, 2002, 113-120.

[18] Salton, G., Singhal, A., Mitra, M., and Buckley, C. "Automatic Text Structuring and Summarization," in Advances in Automatic Text Summarization, The MIT Press, 1999.

[19] Allan, J., Gupta, R., and Khandelwal, V. "Temporal Summaries of News Topic," in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001, 10-18.

[20] Erkan, G., and Radev, D. R. "LexRank: Graph-based Centrality as Salience in Text Summarization," in Journal of Artificial Intelligence Research, Volume 22, 2004, 457-479.

[21] Mihalcea, R., and Tarau, P. A Language Independent Algorithm for Single and Multiple Document Summarization," in proceedings of the International joint conference on natural language processing, 2005, 19-24.

[22] Kleinberg, J. "Authoritative Sources in a Hyperlinked Environment," in Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms, 1998, 668-677.

[23] Brin, S., and Page, L. "The Anatomy of A Large-Scale Hyper-Textual Web Search Engine," in computer networks and ISDN systems archive, vol. 30, issue 1-7, 1998, 107-117.

[24] Sun, J-T., Shen, D., Zeng, H-J., Yang, Q., Lu, Y., and Chen, Z. "Web-Page Summarization Using Clickthrough Data," in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005, 194-201.

[25] Nenkova, A., Vanderwende, L., and Mckeown, K. "A Compositional Context Sensitive Multi-document Summarizer: Exploring the Factors That Influence Summarization," in proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, 2006, 573-580.

[26] Kleinberg, J. "Bursty and Hierarchical Structure in Streams," in Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, 2002, 91-101.

[27] Nallapati, R., Feng, A., Peng, F., and Allan, J. "Event Threading within News Topics," in Proceedings of the thirteenth ACM international conference on Information and knowledge management, 2004. 446-453.

[28] Yang, C. C., and Shi, X. "Discovering Event Evolution Graphs from Newswires," in Proceedings of the 15th international conference on World Wide Web, 2006. 945-946.

[29] Feng, A., and Allan J. "Finding and Linking Incidents in News," in proceedings of the sixteenth ACM conference on conference on information and knowledge management, 2007, 821-830.

[30] Swan, R., and Allan, J. "Automatic Generation of Overview Timelines," in Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000, 49-56.

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

[32] Winston, W.L. "Operations Research," Thomson, 2004.

[33] Hofmann, T. "Probabilistic latent semantic indexing," in Proceedings of the 22th annual international ACM SIGIR conference on Research and development in information retrieval, 1999, 50-57.

[34] Nicholas, C., and Dahlberg, R., "Spotting Topics with the Singular Value Decomposition," Lecture Notes in Computer Science, Vol. 1481, 1998, 82-91.

[35] Rabiner, L. R., and Sambur, M. R. "An Algorithm for Determining the Endpoints for Isolated Utterances," in the Bell System Technical Journal, Vol. 54, No. 2, Feb. 1975, pp 297-315.

[36] Rabiner, L. R., and Schafer, R. W. "Digital Processing of Speech Signals," Prentice-Hall, 1978.

[37] Chen, C.C., and Chen, M.C. "TSCAN: A Novel Method for Topic Summarization and Content Anatomy," in proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, 2008, 579-586.

[38] Lin, C. Y., and Hovy, E. "Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics," in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, 2003, 71-78.

[39] Nenkova, A. "Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference," in Proceedings of the 20th National Conference on Artificial Intelligence (AAAI), 2005, 1436-1441.

[40] Ruan, J. and Zhang, W. "An Efficient Spectral Algorithm for Network Community Discovery and Its Applications to Biological and Social Networks," in Proceedings of the Seventh IEEE International Conference on Data Mining, 2007, 643-648.

Chien Chin Chen received his Ph.D. degree in Electrical Engineering from National Taiwan University, Taiwan, in 2007. He is currently an assistant professor of the department of Information Management at National Taiwan University. His current research interests include text mining, information retrieval, knowledge discovery, and data mining.

Meng Chang Chen received the PhD degree in computer science from the University of California, Los Angeles, in 1989. From 1992 to 1993, he was an associate professor at the National Sun Yat-Sen University, Taiwan. Since then, he has been with the Institute of Information Science, Academia Sinica, Taiwan, where he is currently a research fellow and deputy director. His current research interests include information retrieval, knowledge management and engineering, crowd intelligence, wireless network, and QoS networking.