# Optimizing Temporal Topic Segmentation for Intelligent Text Visualization

**Shimei Pan**[1] **Michelle Zhou**[2] **Yangqiu Song**[3] **Weihong Qian**[4] **Fei Wang**[1] **Shixia Liu**[5]

[1]IBM Research-T.J.Watson, USA
{shimei, fwang}@us.ibm.com

[2]IBM Research Almaden, USA
mzhou@us.ibm.com

[3]HKUST, Hong Kong
yqsong@gmail.com

[4]IBM Research China, China
qianwh@cn.ibm.com

[5]Microsoft Research Asia, China
shliu@microsoft.com

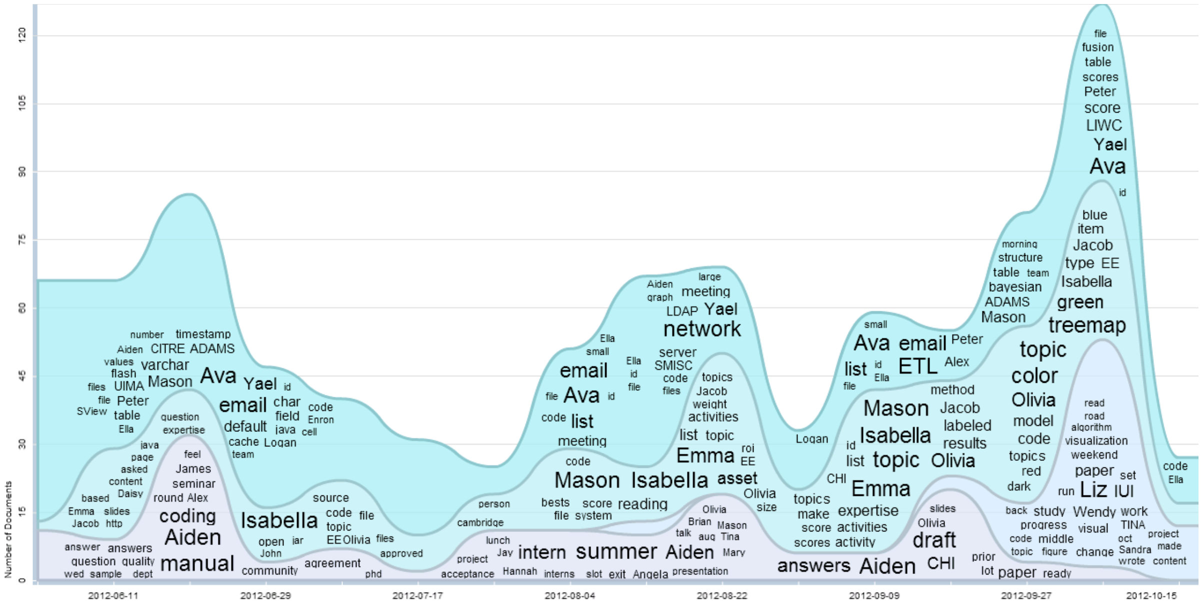

**Figure 1. TIARA-generated visual text summary showing four topic layers.**

## ABSTRACT

We are building a topic-based, interactive visual analytic tool that aids users in analyzing large collections of text. To help users quickly discover content evolution and significant content transitions within a topic over time, here we present a novel, constraint-based approach to temporal topic segmentation. Our solution splits a discovered topic into multiple linear, non-overlapping sub-topics along a timeline by satisfying a diverse set of semantic, temporal, and visualization constraints simultaneously. For each derived subtopic, our solution also automatically selects a set of representative keywords to summarize the main content of the sub-topic. Our extensive evaluation, including a crowd-sourced user study, demonstrates the effectiveness of our method over an existing baseline.

*IUI'13,* March 19–22, 2013, Santa Monica, CA USA.
Copyright © 2013 ACM 978-1-4503-1965-2/13/03...$15.00.

### Author Keywords

Text visualization, topic-based, constrained clustering.

### ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous. [**Artificial Intelligence**]: Natural language Processing – *text analysis.*

## INTRODUCTION

To aid users in effectively gleaning insights and locating critical information from large collections of text documents, we have developed a topic-based, interactive visual analysis tool, TIARA [22]. TIARA provides two main functions. First, it *automatically* generates a time-based visual text summary that conveys key topics derived from a large collection of text. Second, it supports a rich set of interactions that allow users to further explore the created visual summary and examine the text collection from multiple aspects. Figure 1 is an output of TIARA, where each colored layer represents an automatically derived topic. Each layer is depicted by a sequence of word clouds, summarizing the topic content at each time interval. At each time

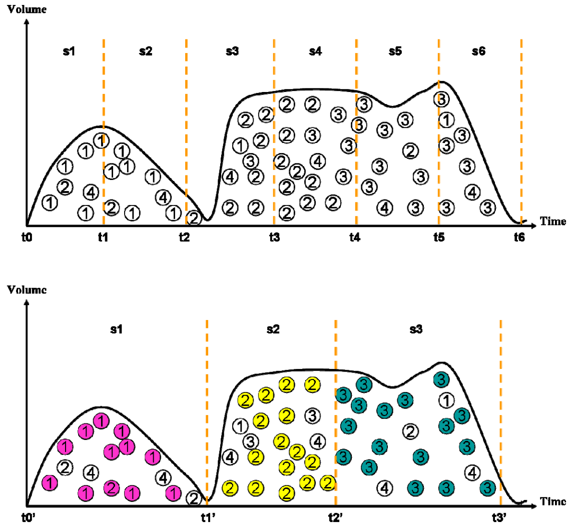

**Figure 2. (a) Top: segmented (s1-s6) by fixed time intervals (t1-t6); and (b) Bottom: optimized segmentation (s1-s3). Each numbered circle denotes a document in a specific sub-topic.**

point, the height of a layer encodes the "strength" of the topic—the number of documents covering the topic at that time. A user can also interact with this visualization, e.g., retrieving text snippets and a full document [22].

To automatically derive the topics shown above, TIARA performs two key tasks: topic extraction and temporal topic segmentation. *Topic extraction* is to automatically derive a set of topics (e.g., sports and politics) from a text collection (e.g., news articles). *Temporal topic segmentation* is to break an extracted topic into a set of sub-topics along the time line so that a user can easily track the temporal evolution of the topic. There are many research efforts on topic extraction, including clustering [16, 32] and probabilistic topic modeling [7, 17, 24]. Currently, TIARA uses the Latent Dirichlet Allocation (LDA) model [7] to automatically extract a set of topics. Although TIARA used simple heuristics to the second task, there is little in-depth research in this area. In this paper, we propose a systematic approach to temporal topic segmentation.

*Temporal topic segmentation* is to split a continuous topic into a sequence of sub-topics over time. It further consists of two steps. First, it must identify a set of meaningful, time-based, semantic transitions to split a topic into multiple, linear non-overlapping temporal segments. Second, it must extract a set of representative keywords to summarize the content of each segment. Both steps are non-trivial, since the system must satisfy a diverse set of constraints, including semantic, temporal, and visual constraints.

**Constraint 1**: *Capturing significant topic transitions*. One of our goals is to help users detect how a topic evolves over time, in particular, identify whether and how the content of a topic has shifted and when. Existing time-based visual text summaries split a topic into multiple temporal segments by a fixed time interval (e.g., monthly) [22]. In reality, however, semantic transitions of a topic may not occur in pre-determined, fixed time intervals (Figure 2a). In such cases, chopping up a topic "unnaturally" along fixed time points not only would miss the natural topic transitions, but may also result in visual redundancies. For example, segments S1 and S2 contain semantically similar content (Figure 2a). To capture natural topic transitions, a temporal segment boundary should be placed near where a significant topic shift occurs (e.g., t2 in Figure 2a)

**Constraint 2**: *Identifying sub-topics with temporal locality.* To effectively visualize topic transitions over time, we are interested in discovering sub-topics that exhibit temporal locality. *Temporal locality* of a topic is a time interval where the documents covering the topic appear most prominently. As shown in Figure 2(a), for example, sub-topic 4 does not have temporal locality, since the documents on this topic are distributed over the entire time span. In contrast, the temporal locality of sub-topic 1 is time interval t0–t2.

**Constraint 3**: *Respecting visual segment boundaries.* As shown in Figure 1, word clouds representing topic segments are placed in a topic layer. To best use the available space in each topic layer, a spatial layout algorithm is often used to place the keywords in the formed peaks and valleys [22]. However, the temporal localities of sub-topics may not be aligned with these visual peaks and valleys. For example, the preferred temporal boundary for sub-topic 1 is t0–t2 in Figure 2(a), which does not line up with the visual boundary t0'–t1' in (b). If we do not respect natural visual shapes, the space use may not be optimal (e.g., placing the keywords for sub-topic 1 in t0–t2 versus t0'–t1'). Moreover, placing a word cloud representing the same topic *across* visual boundaries will disrupt the visual flow. For example, sub-topic 2 spans t2–t4, across the natural visual boundary at t1'. Ideally, temporal and visual boundaries should be aligned to produce a semantically coherent *and* visually pleasing display. Using the above examples, segmenting the topic at t1' is much preferable than at t2.

To address the challenges described above, we develop an optimization-based approach to temporal topic segmentation, including the selection of representative keywords for summarizing each identified temporal segment. Our approach systematically incorporates a set of constraints, including semantic, temporal, and visual constraints. It then satisfies them simultaneously to optimize topic segmentation and keyword selection. To demonstrate the effectiveness of our work, we have conducted both algorithmic and end user evaluations. Our results show significant improvements over an established baseline. As a result, our work offers two unique contributions:

1. Our optimization-based approach to temporal topic segmentation is effective, as it satisfies a diverse set of semantic, temporal, and visualization constraints for creating a time-based, visual text summary.

2. Our constraint-based framework is extensible, since it can easily incorporate additional constraints when needed (e.g., geo-spatial constraints to group together segments originated from the same geolocation).

In the rest of the paper, we first summarize related work before giving an overview of our tool. We then explain our approach in detail on temporal topic segmentation, followed by the evaluations. We also discuss limitations of our work and potential improvements before concluding.

## RELATED WORK

Similar to our work, document segmentation [5, 10, 19, 20, 23, 29] is to identify a set of topic transitions in a single or multiple document(s). There are a number of approaches to single document segmentation, such as detecting segmentation points by specific speech or lexical cues [29], identifying lexical or semantic changes between adjacent text blocks [19, 20], and probabilistic topic segmentation [10, 23]. On the other hand, multi-document segmentation takes multiple *similar* documents (e.g., multiple news articles from different sources reporting the same event) to find a set of sub-topics on the shared topic [21, 28]. Compared to the existing work on document segmentation, ours focuses on *temporal* segmentation of a set of heterogeneous documents, which has not been addressed before. Moreover, we must consider *visual* constraints during topic segmentation to ensure an effective temporal visualization of the derived-segments.

Similar to our goal of investigating topics along time, there is a rich body of work on temporal topic analysis [2, 8, 12, 30, 31, 33]. The main purpose of these works is to incorporate time as an *input* to improve topic discovery results. For example, one effort is to discover how the topics in one year evolve from those of the previous year [8]. In such a model, documents are first grouped by a *fixed* time frame (e.g., per year). A topic model is then used to derive multiple topics from the documents in each time frame. In contrast, our work is to identify and *output* proper time frames that best segment a topic by satisfying a set of semantic, temporal, and visualization constraints.

Our work is also related to existing efforts in creating topic-based, interactive visual text summarization systems [11, 15, 22]. These works focus on developing visual metaphors and interactions that allow users to understand and explore the derived topics. To generate a time-based, visual summary, these systems also need to segment topics along a timeline. However, existing approaches are often ad hoc. For example, in our earlier work on topic segmentation, the temporal segment boundaries are mainly determined by visual constraints, such as the peaks and valleys of a topic layer [22]. If a topic layer has too few natural visual segment boundaries (the shape of the topic layer is mostly flat), additional segment boundaries are then introduced using fixed time intervals. Nonetheless, such approaches cannot guarantee that the derived temporal segments capture the correct significant topic transitions. In contrast, here we present a systematic approach to temporal topic segmentation by balancing a diverse set of constraints.

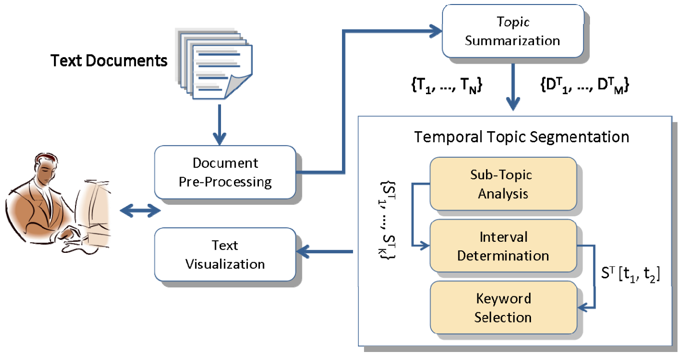

**Figure 3. System overview.**

## SYSTEM OVERVIEW

Figure 3 provides an overview of TIARA. The input to TIARA is a collection of text documents. The *document pre-processing* component first extracts the title and the main body text (e.g., the body of a news story) and related meta data (e.g., the author and time information) from each document. The extracted information is then sent to the *topic summarization* component, which now uses an LDA-based approach to automatically extract a set of topics. Each derived topic is associated with a set of documents.

Given a derived topic (T) and a set of documents covering the topic $\{D^T_1, \ldots, D^T_M\}$, the *sub-topic analysis* component identifies a set of sub-topics $\{S^T_1, \ldots, S^T_K\}$ that satisfy a set of semantic, temporal, and visualization constraints. The *interval determination* component computes the temporal boundary for each sub-topic $S^T$ [t1, t2]. Finally, the *keyword selection* component identifies a set of keywords to represent each derived sub-topic. Given the derived topics and sub-topics, the *text visualization* component generates a time-based, interactive visual text summary. Our focus below is on the three components highlighted in red.

## OUR APPROACH

We describe our approach in two parts: (1) we describe a three-step process that uses constraint-based topic analysis to identify a set of sub-topics of a given topic in a linear, non-overlapping temporal sequence; and (2) we present a keyword selection method, which selects a set of representative keywords to summarize each derived sub-topic.

### Temporal Segmentation

There are three main steps in temporal segmentation: (1) constraint generation, (2) topic analysis with the generated constraints, and (3) temporal segment boundary determination based on the topic analysis results.

#### Step 1: Generating Visual and Temporal Constraints

As mentioned earlier, our goal is to identify temporal subtopics that satisfy a set of semantic, temporal, and visual constraints. Since typical topic analysis already incorporates latent semantic constraints [7, 27, 32], there is no need to encode semantic constraints explicitly. We thus need only encode temporal and visual constraints. To incorporate these constraints into a constraint-based topic modeling process, we represent them as must-links and cannot-links that enforce specific relationships between two documents [6, 27, 32]. More specifically, *must-link* ensures that related documents be in the same sub-topic (e.g., documents A and B should be in the same sub-topic), while *cannot-links* indicate that related documents should *not* be associated with the same sub-topic (e.g., documents C and D should be in different sub-topics).

Given a derived topic and a set of documents *D* relevant to the topic, we add a must-link or a cannot-link between each document pair $d_i$ and $d_j$ in *D* as follows:

(1) Adding a *cannot-link* constraint between $d_i$ and $d_j$ if they fall in different visual segments. Here visual segment boundaries are computed by our visualization component prior to topic segmentation. Since each document is associated with a time stamp, we use its time stamp to determine which visual segment the document falls in. Whenever possible, our goal is to place the keywords representing a sub-topic in a single visual segment instead of across the boundaries of two visual segments. To achieve this goal, the added cannot-links discourage documents falling in different visual segments to form a single sub-topic.

(2) If the temporal distance between the two documents is within a threshold $\partial$ (e.g., within a week), add a *must-link* between them. These must-links encourage documents that are temporally close to each other to be in the same sub-topic. As a result, these constraints help discover the temporal locality of sub-topics.

(3) Otherwise, if the temporal distance between the two documents is above a threshold $\Delta$ (e.g., above 6 months), add a *cannot-link* between them. These cannot-links will prevent temporally distant documents from being placed in the same sub-topic. In other words, these links discourage the formation of sub-topics that do not have temporal locality.

Here, thresholds $\partial$ and $\Delta$ are used to define the temporal proximity of two documents. Currently, both values are set empirically based on the properties of the target dataset (e.g., the time span and the distribution of documents).

Note that, multiple constraints may co-exist between a pair of documents. For example, two documents fall in two visual segments, so a cannot-link will be added in step (1). Since they are also temporally close enough, a must-link will also be added in step (2). In such cases, we must assign priority to different constraints. Currently, the priority is given to visual constraints. Our rationale is that the number of visual constraints is far fewer than the number of the temporal constraints. Our heuristic promotes more specific visual constraints and prevents them from being overpowered by more general temporal constraints.

After the must-links and cannot-links are generated, we employ constraint-based topic analysis to derive sub-topics that satisfy these constraints.

**Step 2: Performing Constraint-based Topic Analysis**

To incorporate the visual and temporal constraints specified above in a topic modeling process, we employ constraint-based topic analysis. There are two families of constraint-based topic analysis: (a) Constrained LDA [3, 4, 24] and (b) constrained clustering [6, 27, 32]. Since constrained LDA is a relatively new effort, without extensions, none of the existing methods can encode the temporal and visual constraints that must be satisfied in our application. We thus decide to use constrained clustering.

Constrained clustering is a class of semi-supervised clustering algorithms that allows a system to incorporate additional user or application constraints during clustering. Frequently, these constraints are soft constraints and not all of them will be satisfied in the final solution. In general, constrained clustering is formulated as constraint-based optimization where its objective function captures both cluster coherence and additional constraints. This property fits our goal quite well, since it ensures the semantic coherence of discovered sub-topics while satisfying additional temporal and visual constraints.

Among all the constrained clustering algorithms, we adopt *constrained co-clustering* for two reasons. First, it is flexible as it can cluster multiple variables simultaneously. Traditional constrained clustering methods that cluster one random variable at a time are special cases of constrained co-clustering methods. Second, it is more effective in text analysis than one dimensional clustering, since it can cluster documents and words jointly [16]. To leverage the state-of-the-art constrained co-clustering methods, we decide to experiment with two top-performed algorithms but with very different clustering frameworks: Constrained Information-Theoretic Co-Clustering (CITCC) [27] and Constrained Co-Clustering with Non-negative Matrix Tri-Factorization (CCCTriNMF) [32].

*CITCC* combines the benefits of information-theoretic co-clustering with constrained clustering for textual documents [27]. It utilizes a two-sided hidden Markov random field to model both the document and word constraints. In *CITCC*, the clustering process is formulated as constraint-based optimization with the following objective function:

$$\arg\min D_{KL}(p(D,V,\hat{D},\hat{V} \| q(D,V,\hat{D},\hat{V}) \\ +\sum_{d_{m_1}}^{M}\sum_{d_{m_2}\in M_{d_{m_1}}} E(d_{m_1},d_{m_2}\in M_{d_{m_1}}) \\ +\sum_{d_{m_1}}^{M}\sum_{d_{m_2}\in C_{d_{m_1}}} E(d_{m_1},d_{m_2}\in C_{d_{m_1}}) \qquad (1) \\ +\sum_{v_{i_1}}^{V}\sum_{v_{i_2}\in M_{v_{i_1}}} E(v_{i_1},v_{i_2}\in M_{v_{i_1}}) \\ +\sum_{v_{i_1}}^{V}\sum_{v_{i_2}\in C_{v_{i_1}}} E(v_{i_1},v_{i_2}\in C_{v_{i_1}})$$

where $D=\{d_1,..,d_M\}$ and $V=\{v_1,..,v_V\}$ are document and word sets, $\widehat{D}$ and $\widehat{V}$ are the document and word clusters, $M_{d_m}$ and $C_{d_m}$ are the must-links and cannot-links on the latent document cluster label $L_{d_m}$ for document $d_m$. Similarly, $M_{v_i}$ and $C_{v_i}$ are the must-links and cannot-links on the latent word cluster label $L_{v_i}$ for word $v_i$; $E$ is the energy function for must-links and cannot-links.

In contrast, *CCCTriNMF* [32] is matrix-factorization-based, constrained co-clustering. Similar to CITCC, the constraints are encoded as must-links and cannot-links. It finds a solution to optimize the following objective:

$$\underset{G_1\geq 0, G_2\geq 0}{Min}\left\|R_{12}-G_1SG_2^T\right\|^2+Tr(G_1^T\Theta_1G_1)+Tr(G_2^T\Theta_2G_2) \qquad (2)$$

where $R_{12}$ is a co-occurrence document and word matrix, $G_1$ and $G_2$ are the cluster indicator matrices for document and word. $Tr(G_1^T\Theta_1G_1)$ and $Tr(G_2^T\Theta_2G_2)$ are penalties for violating the must-link and cannot-link constraints of documents and words, and $\Theta_1$ and $\Theta_2$ are penalty matrices.

**Step 3: Determining Temporal Boundaries**

Although Step 2 derives a set of document clusters representing a set of sub-topics, the clusters have not yet been assigned temporal boundaries. In Step 3, we employ two methods to determine the temporal boundaries of derived sub-topics: (1) by cluster centers and radiuses, and (2) by smoothed labels.

***Cluster Center and Radius-based (CR)*** Based on the constrained clustering results, we extract the *temporal center* and the *temporal radius* for each derived document cluster (sub-topic). The *temporal center* ***c*** of a document cluster is defined as the median time stamp after sorting all the documents in the cluster on the time line, half of them is before ***c*** and half is after ***c***. The *temporal radius* ***r*** of a cluster is used to define a range [***c-r, c+r***] so that *x%* of the documents in the cluster are within that range. Normally, *x* is set empirically based on the target data set. Figure 4 shows the cluster centers and radiuses of two document-clusters.

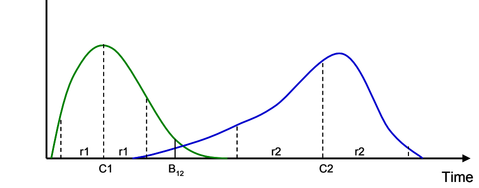

**Figure 4. Cluster center and radius-based segmentation.**

To associate the clusters with time, we sort all the clusters on the time line based on their temporal centers. The segment boundary between two adjacent clusters $c_1$ and $c_2$ is determined based on their cluster radiuses. Given distance $d$ between the centers of two adjacent clusters, the segment boundary $b_{1,2}$ is decided by:

$$b_{1,2} = c_1 + \frac{r_1}{r_1 + r_2} \times d \qquad (3)$$

Since the size of the derived clusters can be imbalanced (e.g., some clusters are with thousands of documents, while others may only have a few), we performed smoothing to remove clusters that are "too small to visualize effectively". Here, we consider a cluster whose size is less than 10% of the average cluster size "too small to visualize".

***Smoothed Label-based (SL).*** Our second approach is to use a sliding window to determine temporal boundaries of the derived clusters (sub-topics) [5, 13, 20]. We first directly sort all the documents on the timeline based on their time stamps. We then use a sliding window of size *N* to examine *N* documents at a time and identify those whose cluster labels do not conform to the main cluster where the majority lies. Such a document is then re-assigned to the main cluster within the window. Based on the smoothed document labels, a temporal boundary occurs whenever there is a change of cluster labels between two adjacent documents on the time line. In our experiments, *N* was set empirically based on the target dataset. In general, larger window size will result in smaller number of segments. Similar to the *CR*-based algorithm, "small clusters" may be removed during *SL* segmentation due to "smoothing".

### Selection of Representative Segment Keywords

After Step 3, each document cluster (sub-topic) is now associated with a temporal boundary, which is called a *temporal segment*. For each temporal segment *S*, we select a set of topic keywords to best summarize the documents in the segment. The method for selecting topic keywords varies depending on which clustering algorithm is used.

For CITCC, the rank of a word $v_i$ is computed using:

$$r_{v_i \in S} = \sum_{k_v} q(\widehat{d}_{k_d}, \widehat{v}_{k_v}) q(v_i \mid \widehat{v}_{k_v}) \qquad (4)$$

For each word $v_i$ in any of the documents in segment *S,* we compute the rank of the word $r_{v_i}$ by formula (4) where $\widehat{d}_{k_d}$ is the document cluster associated with segment *S*, $q(\widehat{d}_{k_d}, \widehat{v}_{k_v})$ is the probability of a word cluster $\widehat{v}_{k_v}$, $k_v$ =1, 2, 3. …, *K*, be associated with the document cluster $\widehat{d}_{k_d}$, and $q(v_i \mid \widehat{v}_{k_v})$ is the probability a word $v_i$ is associated with the word cluster $\widehat{v}_{k_v}$.

To improve efficiency, here we only focus on the *most relevant word clusters*. The relevant word clusters for a document cluster $\widehat{d}_{k_d}$ are defined as those whose $q(\widehat{d}_{k_d}, \widehat{v}_{k_v})$ is above a certain threshold.

For CCCTriNMF, the rank of a word $v_i$ is computed using:

$$r_{v_i} = [S.G_2^T]_{k,i} \quad (5)$$

where *S* is the cluster center matrix in equation (2), $G_2^T$ is the transpose of the word cluster indicator matrix $G_2$, *k* is the document cluster label associated with the current segment, and *i* is the word index of $v_i$.

Based on the word ranks defined in equations (4) and (5), we retrieve the top-*K* keywords to represent each segment.

## ALGORITHMIC EVALUATION

To test the performance of our approach, we first conducted a set of experiments to measure the performance of our algorithms in different settings.

### Data Sets

We used two data sets in our experiments. The first data set is a collection of 7000+ emails over the course of two years. The second is a collection of 13,000+ New York Times articles, published during the time period of six months.

### Settings and Baseline

TIARA first ran LDA topic modeling on each of our data set to derive *N* topics. For emails, *N* was set to 10. For news articles, *N* was set to 30.

For each of the *N* topics, we extracted all the documents relevant to the topic. A document *k* is considered relevant to a topic *i* if its LDA-derived document topic probability $\theta_{k,i}$ is above a threshold; or if $\theta_{k,i}$ is the highest among all the $\theta_{k,j\ (j \neq i)}$ if none is above the threshold. In our experiments, the threshold was set to 0.4 for emails and 0.3 for news articles. For each LDA-derived topic, we then ran our temporal segmentation algorithms. We repeated the same process 10 times, each with a different random seed for CITCC and CCCTriNMF. All the results presented here are the averages over 10 random runs across all the *N* topics.

For comparison, we used the previous version of TIARA as a baseline [22]. This is the first and only method known to us that performs temporal topic segmentation for text visualization. The temporal boundaries in the baseline were determined first by visual segment boundaries, augmented with additional segments based on fixed time intervals to make the total number of segments the same as the document cluster number used in CITCC and CCCTriNMF. To select segment keywords, the baseline computed the rank of a keyword in each temporal segment based on its segment-specific *tf*idf* scores. Here, *tf* is the accumulative frequency of a word in all the documents in a segment. *idf* is computed the same way as that in typical IR tasks. We used the top-*K* keywords for each segment.

### Metrics

We used three objective metrics to evaluate our algorithms.

*Topic Completeness* measures how relevant the selected segment keywords are to the current topic. As shown in Figure 1, the semantics of a topic is defined by the keywords associated with all the segments in the topic. Thus, one critical role played by these keywords is to help users understand the semantics of a topic. A high score in topic completeness indicates that it is easy for a user to grasp the overall semantics of a derived topic through the keywords in all temporal segments. In contrast, a low score implies that the extracted segment keywords may not be a good representation of the topic.

**Table 1. Experimental results on two data sets. Except for visual alignment, all the numbers in bold represent significant improvement over the baseline. The bold numbers with * in visual alignment represent significant difference between the two segmentation algorithms under the same clustering condition.**

| Method | Topic Completeness | | Topic Distinctiveness | | Visual Alignment | |
|---|---|---|---|---|---|---|
| | **Email** | **NYT** | **Email** | **NYT** | **Email** | **NYT** |
| Baseline | .260 | .426 | .103 | .005 | .000 | .000 |
| CITCC-CR | **.561** | **.575** | .079 | **.111** | .082 | .050 |
| CITCC-SL | **.561** | **.573** | .077 | **.118** | **.067*** | **.043*** |
| CCCTriNMF-CR | **.440** | **.494** | **.128** | **.147** | .097 | .066 |
| CCCTriNMF-SL | **.432** | **.479** | **.115** | **.156** | .095 | .068 |

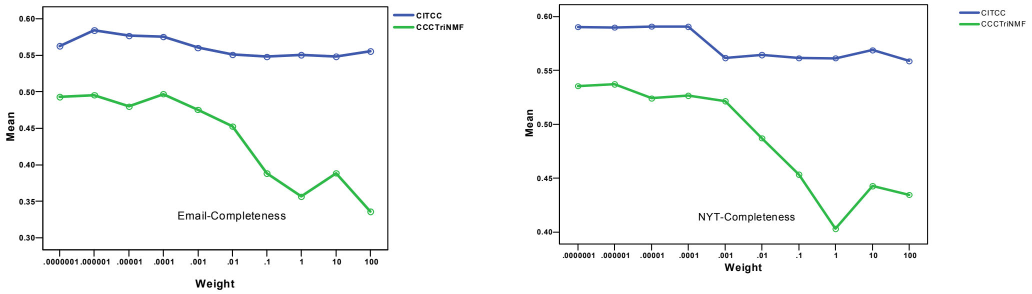

**Figure 5. The impact of constraint weight on topic completeness.**

To compute topic completeness, for each topic, we compare the union of all the keywords in all the derived segments (sub-topics) with the original top 50 keywords derived by LDA for that topic. Based on the number of overlapping keywords, we compute *precision* to measure among all the keywords in the segment keyword union, the percentage of which also belongs to the top 50 topic keywords derived by LDA. We also compute *recall* to measure among all the top 50 keywords from LDA, the percentage of which also appears in the union of all the segment keywords. Finally, we compute the *F-measure,* the harmonic mean of precision and recall*,* as the overall assessment for topic completeness. The completeness score in Table 1 is the average *F-measure* over all the topics over 10 random runs.

*Topic Distinctiveness* measures how one segment differs from another based on their associated keywords. It helps us evaluate how well the temporal segments produced by our algorithm capture significant topic transitions. Here, we use the average pair-wise Jensen-Shannon (JS) divergence to measure the distinctiveness of two sets of keywords. Suppose that a topic is segmented into L parts. Each segment $l$ has a normalized keyword histogram $h_l$ s.t. $\sum_m h_l(m)=1$ where $m$ is the word index in the vocabulary.

Given two normalized keyword histograms $h_i$ and $h_j$, first we compute the Kullback-Leibler (*KL*) divergence:

$$D_{KL}(h_i \| h_j) = \sum_{k=1}^{v} h_i(k) \log \frac{h_i(k)}{h_j(k)} \quad (6)$$

Then we compute the Jensen-Shannon divergence:

$$D_{JS}(h_i \| h_j) = \frac{1}{2} D_{KL}(h_i \| \bar{h}) + \frac{1}{2} D_{KL}(h_j \| \bar{h}) \quad (7)$$

where $\bar{h} = 1/2(h_i + h_j)$ . The final distinctiveness score of a topic is the average pair-wise JS score over all the segments in a topic. After deriving the topic distinctiveness score for each topic, we then compute the average over all the topics over 10 random runs.

*Visual Segment Alignment* measures how well the derived temporal segments align with visual segments. While our first two metrics evaluate the semantic coherence of derived temporal segments, this metric evaluates how well our algorithm satisfies the visualization constraints. Here the visual segment boundaries were pre-computed by our system based on the shape of each topic layer. We computed the average distance between each visual segment boundary and its nearest temporal segment boundary inferred by our algorithms. We normalized the distance so that it is in the range of [0, 1]. Here, the lower the score (distance) is, the better the alignment is. The scores shown in Table 1 are the averages over all the visual segment boundaries for all the topics over 10 random runs.

### Results and Analysis

As shown in Figure 5 for *topic completeness*, all our algorithms outperformed the baseline on both data sets. The difference is statistically significant based on paired *t*-test ($p<0.001$). This result strongly suggests that by employing constrained clustering, our algorithm is much more capable of identifying topic-related keywords. Low topic completeness score often means it is hard for users to grasp the semantics of a topic based on the sum of the segment keywords. Moreover, between the two clustering algorithms, CITCC-based methods performed better than CCCTriNMF-based methods for topic completeness ($p<0.001$).

For *topic distinctiveness*, CCCTriNMF-based approaches performed the best on both data sets. They performed significantly better than the baseline ($p<0.001$). They also performed significantly better than the CITCC-based methods ($p<0.001$). Note that the topic distinctiveness scores for the baseline system in both data sets are quite low (**Error! Reference source not found.**). Since the baseline used only visual boundaries and fixed time intervals to do segmentation, this may cause redundancies by splitting similar content into multiple segments (e.g., S1 and S2 in Figure 2(a)).

For *visual segment alignment*, since by definition, the alignment score for the baseline is always 0, we did not compare our results with the baseline. Instead, we compared the performances between our two temporal alignment algorithms (i.e., *CR* versus *SL*) under the same clustering condition. Based on paired-t test, if used in combination with CITCC, the SL-based algorithm performed significantly better than the CR-based method on both datasets ($p<0.001$). There is no statistically significant difference

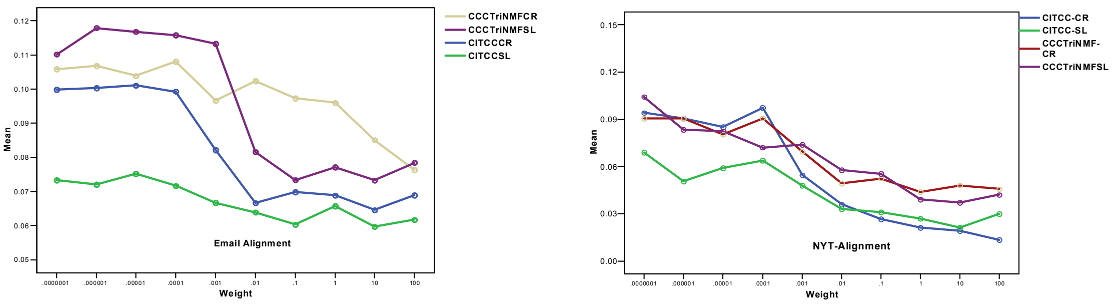

**Figure 6: The impact of constraint weight on visual alignment.**

between these two approaches, if they were used in combination with CCCTriNMF.

Since both CITCC and CCCTriNMF attempt to strike a balance between discovering coherent sub-topics and satisfying additional visual and temporal constraints, we also investigated how constraint weight affects the results. In this experiment, we varied the weights of must-links and cannot-links in CITCC and CCCTriNMF from 0.0000001, the lowest, to 100, the highest.

Figure 5 shows the impact of weight on topic completeness on both datasets. As shown in Figure 5, the topic completeness is mainly determined by the constrained clustering algorithms (i.e., CITCC or CCCTriNMF) and not sensitive to the two temporal segmenting algorithms (CR or SL). We thus show the average topic completeness for each clustering method, regardless of the segmentation algorithm used. As shown in Figure 5, when the weight of constraints increases, the topic completeness for CCCTriNMF deteriorates significantly. In contrast, CITCC holds its performance much better than CCCTriNMF.

Similarly in Figure 7, we show how constraint weight affects topic distinctiveness. Here, in CCCTriNMF (the green lines), when the constraint weight increases, the distinctiveness of the segments increases slightly. In contrast, for CITCC (the blue lines), increasing constraint weights seems having little impact on email data and negative impact on the news data on the topic distinctiveness.

Moreover, the impact of constraint weight can be clearly observed in Figure 6, which shows how visual alignment scores change with the constraint weights on two data sets. Each curve in these graphs represents a combination of clustering (CITCC and CCCTriNMF) and temporal segmentation algorithm (CR and SL). As shown in these graphics, when the weight increases, the alignment score decreases, indicating a better visual-temporal boundary alignment. This trend holds regardless of the data set, the clustering algorithm used, and the temporal alignment algorithm used. These results are significant and demonstrate the effectiveness of our work in satisfying visualization constraints. Moreover, among the four combinations that we have investigated, CITCC-based algorithms outperformed CCCTriNMF-based ones; and the combination of CITCC_SL performed the best in most situations.

## USER STUDY

In addition to our algorithmic evaluation, we also designed and conducted a crowdsourced user study to evaluate the effect of our work on user task performance. Specifically, we evaluated how our method aided users in their text analytic tasks against the baseline.

### Data and Method

In this study, we used the email data set, since we need to compare the user task performance against the ground truth that we have already known. Moreover, we focused on testing TIARA's capability in helping users perform practical text analysis tasks. For this purpose, TIARA created two visualizations (Figure 8). Figure 8(a) shows the temporal segmentation results of a topic produced by the baseline [22]. Figure 8(b) displays the results of our CITCC_SL

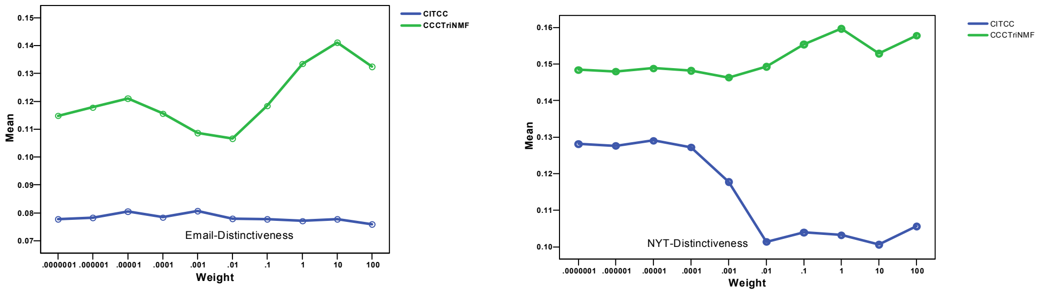

**Figure 7. The impact of constraint weight on topic distinctiveness.**

algorithm, since it outperformed others in most cases by our algorithmic evaluation. The displayed topic involved 155 emails, dated from 6/1/2012 to 10/10/1012. The two visualizations were the same (e.g., size, color, and shape) except the location and content of the keyword clouds.

We designed two surveys, one for each visualization. The surveys contained identical instructions and questions except the visualization. Each participant was first given a brief introduction to the task, including the data and the visualization. S/he was then instructed to view the visualization carefully to answer four multiple-choice questions, each of which is a text analytic task to gain an understanding of the selected topic. The first two questions (Q1 and Q2) were designed to test a user's overall understanding of the derived topic. The next two questions (Q3 and Q4) tested a user's understanding of specified sub-topic(s) and associated temporal information.

Q1: *Who is most involved in Project X in its entire course?*

Q2: *What are the main sub-topics discussed in this project?*

Q3: *When did event E occur*?

Q4: *Which event(s) occurred during the time frame X?*

We used a between-subject design and sent 50 assignments for each survey on CrowdFlower [1], a crowdsourcing platform. To ensure the quality, we recruited workers located in the U.S. only and also limited a worker to work on only one of the two surveys once. We collected 51 responses for the baseline and 54 responses for our method.

### Study Results and Analysis

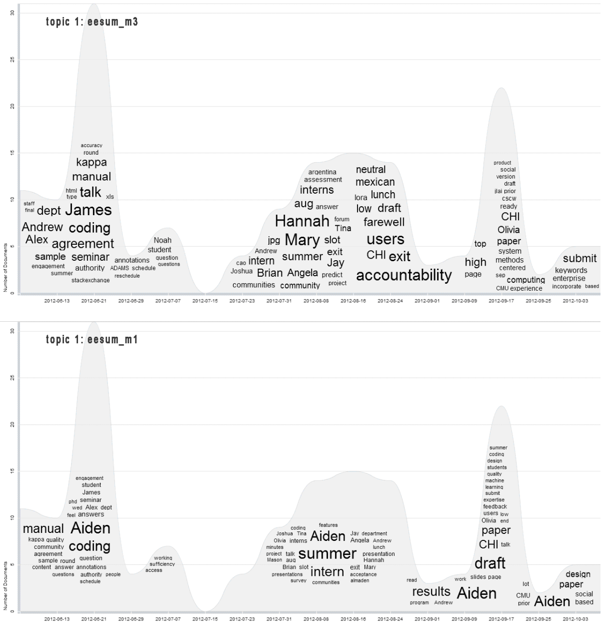

**Figure 8. Two visualizations used in our user study. (a) Top: baseline, and (b) Bottom: created by our algorithm.**

From the collected survey results, we examined two metrics. One is *task completion time*, which was computed using the job start and submission time automatically recorded by CrowdFlower. The other is *task success rate*, measuring the percentage of correct answers that a participant produced. Overall, the participants who used Figure 8(b) took longer (*mean* = 4.8 minutes) to complete their tasks than those with the baseline (*mean* = 4.2 minutes), although the difference was not significant (*p<0.08* based on independent samples *t*-test). However, these participants achieved a much higher task success rate (*mean* = 81% for our method versus 18% for the baseline). The difference is statistically significant (*p<0.001*). This means that although those participants spent a bit more time, they were able to perform the analytic tasks much better using the visualization produced by our method.

We further examined the effect of different tasks on the task success rate since our survey included two types of analytic tasks: (1) achieving an overall understanding of a topic (Q1 and Q2), and (2) extracting specific aspects of a derived temporal sub-topic (Q3 and Q4). We found that the difference in task success rate was even greater in Type 1 task, 83% vs. 6%, while smaller, 79% vs. 30%, for Type 2 task.

We believe the difference is because our topic segmentation approach models latent semantics of sub-topics and ensures that extracted temporal segments are semantically coherent and distinctive (consistent with our *topic completeness* and *topic distinctiveness* evaluation). In contrast, the baseline relies on only a *tf***idf* based model, and important concepts (especially those appearing often in all the documents) might be discounted without showing up prominently in the display. Moreover, without a coherent topic model, the keywords selected by the baseline were disconnected. In contrast, our method selected keywords that are semantically and temporally meaningful.

As demonstrated by our study, the ability to capture semantic and temporal coherence of sub-topics and their representative keywords is critical to effective visual text analysis. Users often rely on the temporally distributed word clouds to interpret and differentiate sub-topics. In this study, however we did not separately test the effect of visual coherence (e.g., alignment with visual segment boundaries) on user task performance. Further studies are needed to verify the impact of visual coherence on user task performance.

## DISCUSSION

One distinct advantage of our approach is its extensibility, as new constraints can be easily incorporated in a topic analysis process. Here we briefly discuss handling of new types of constraints and limitations in our current work.

### Handling New Types of Constraints

It is quite straightforward for our model to support new types of constraints, such as geo-spatial constraints. Assume that we want to gauge the opinions posted on social media (e.g., Twitter) towards the 2012 US presidential candidates by geographical locations over time. In this case, we want to group the posts not only based on their semantic themes and temporal localities, but also by the authors' geolocations. To achieve this goal, we just need to incorporate must-links and cannot-links that indicate geographical relationships between two documents based on the locations of their authors. TIARA can then perform topic analysis while balancing *all* the constraints, including the geo-spatial constraints. In short, our approach is easily applicable to new situations where new types of constraints need to be incorporated in topic analysis.

### Content-based Modeling of Document Time

To segment a topic temporally, we must obtain the temporal information of each document. Currently, we associate a document with a time stamp, which, in most cases, is the document creation time (e.g., when an email is sent or a news article is published). To capture content-based temporal information, we must then perform temporal information extraction for each document, e.g., inferring the temporal intervals of described events. Although there are multiple temporal reasoning models in natural language processing, finding a suitable tool that can handle a large collection of documents with high accuracy is non-trivial [25].

### Setting Model Parameters

In our current framework, we need to set multiple model parameters, such as word and document cluster number, the size of the sliding window for SL-based segmentation, and weights of various constraints. Although we can use heuristics to estimate the parameters for a data set based on its characteristics (e.g., size of the data set), in reality, the parameters will still need to be tuned empirically for each new dataset. Optimizing such parameters for each dataset *automatically* is an active research topic in Machine Learning and beyond the scope of the paper.

### Topic Analysis with Enriched Constraint Models

Currently, we allow the use of must-links and cannot-links between a pair of documents/words to specify additional constraints in a topic modeling process. In reality, constraints could be far more complex. For example, we may want to specify finer-grained temporal/visual constraints between two documents that a must-link or cannot-link cannot capture [26]. While there are many sophisticated constraint models and solvers that we could leverage, the key challenge is how to incorporate such models with topic analysis that handles complex latent semantic constraints. Even more challenging is how to satisfy all the constraints *efficiently* when millions of documents are involved.

## CONCLUSION

To create an effective, time-based visual summary of text, in this paper, we present a constraint-based topic analysis approach to temporal topic segmentation. Given a topic derived from a set of text documents, our approach automatically splits it into a set of sub-topics (segments) spanning over multiple linear, non-overlapping temporal intervals. It does so by systematically incorporating and simultaneously satisfying a diverse set of constraints, including semantic, temporal, and visualization constraints. As a result, our approach produces semantically coherent temporal segments that capture significant topic transitions. Moreover, the identified temporal boundaries are aligned with the natural visual boundaries of a topic layer to optimize the use of space for displaying each word cloud, summarizing each sub-topic.

We have also conducted extensive experiments to measure the performance of our methods by three objective metrics: topic completeness, topic distinctiveness, and visual segment alignment. Our results have demonstrated the significant advantage of our method over an existing baseline. In particular, our Constrained Information-Theoretic Co-Clustering (CITCC) topic analysis with the Smoothing Label temporal alignment approach performs the best in most cases. Furthermore, our crowdsourced user study demonstrates the effectiveness of our work in aiding users completing practical text analysis tasks.

## REFERENCES

1. http://www.crowdflower.com
2. Alonso, O., Gertz, M. and Baeza-Yates, R. 2009. Clustering and Exploring Search Results using Timeline Constructions. *CIKM'09*, 97-106.
3. Andrzejewski, D., Zhu, X., Craven, M., and Recht, B. 2011. A Framework for Incorporating General Domain Knowledge into Latent Dirichlet Allocation using First-Order Logic. *IJCAI'2011*, 1171-1177.
4. Andrzejewski, D., Zhu, X., and Craven, M. 2009. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors. *ICML*, 4.
5. Banerjee, S. and Rudnicky, A. 2006. A TextTiling Based Approach to Topic Boundary Detection in Meetings. In proceedings of *the Interspeech*. pp 57-60.
6. Basu, S., Bilenko, M. and Mooney, R. 2004. A Probabilistic Framework for Semi-Supervised Clustering. *SIGKDD'04*, 59-68.
7. Blei, D., Ng, A. and Jordan, M. 2003. Latent Dirichlet Allocation. *J. of Mach. Learn. Res.*, 3:993–1022.
8. Blei, D., Lafferty, J. 2006. Dynamic topic models. *ICML'06,* 113–120.
9. Basu, S., Banerjee, A., and Mooney, R. J. 2002. Semi-supervised clustering by seeding. *ICML'02*, 27-34.
10. Brants, T., Chen, F. and Tsochantaridis, I., 2002 Topic-based document segmentation with probabilistic latent semantic analysis, *CIKM' 02,* 211 – 218.
11. Carenini, G., Ng, R. Pauls, A.2008: Interactive multi-media summaries of evaluative text. *IUI'08*, 124-131.

12. Chi, Y., Song, X., Zhou, D., Hino, K. and Tseng, B. 2007. Evolutionary spectral clustering by incorporating temporal smoothness. *SIGKDD'07*, 153-162.

13. Chu, C.-S.J. 1995. Time Series Segmentation: A Sliding Window Approach. *Information Sciences*, 85 (1):147-173.

14. Chuang, J., Ramage, D., Manning, C., and Heer, J. 2012. Interpretation and trust: designing model-driven visualizations for text analysis. *CHI'12*, 443-452.

15. Cui, W., Liu, S., Tan, L., Shi, C., Song, Y., Gao, Z., Qu, H., and Tong, X. Textflow: Towards better understanding of evolving topics in text. *IEEE Trans. Vis. Comput. Graph*. 17, 12 (2011), 2412–2421.

16. Dhillon, I., Mallela, S. and Modha, D. 2003. *Information* Theoretic Co-Clustering. *SIGKDD'03*, 89–98.

17. Dredze, M., Wallach, H., Puller, D., and Pereira, F. 2008. Generating Summary Keywords for Emails Using Topics. *IUI'09*, 199-206.

18. Galley, M., McKeown, K., Fosler-Lussier, E. and Jing., H. 2003. Discourse Segmentation of Multi-party Conversation. *ACL'03*, 562–569.

19. Hearst, M. 1994. Multi-paragraph segmentation of expository text. *ACL'94*, 9-16.

20. Hearst, M. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33-64.

21. Jeong, M. and Titov, I. 2010 Multi-Document Topic Segmentation. *CIKM'2010*, 1119-1128.

22. Liu, S., Zhou, M., Pan, S., Qian, W., Cai, W., Lian, X. 2009. Interactive Topic-based Visual Text Summarization and Analysis, *CIKM'09*, 543-552.

23. Misra, H., Yvon, F., Jose, J. and Cappe, O. 2009. Text segmentation via topic modeling: an analytical study. *CIKM '09*, 1553-1556.

24. Ramage, D., Manning, C. and Dumais, S. 2011. Partially Labeled Topic Models for Interpretable Text Mining. *SIGKDD'11*, 457-465.

25. Sanampudi, S. and Kumari, G. Temporal reasoning in natural language processing: a survey. *Intl. J. of Comp. Apps*. 1(4): 53-57, 2010.

26. Schrier, E., Dontcheva, M., Jacobs, C., Wade, G. and Salesin D. *.IUI '08*, Adaptive layout for dynamically aggregated documents. 99-108.

27. Song, Y., Pan, S., Liu, S., Wei, F., Zhou, M. and Qian, W. 2010. Constrained co-clustering for textual documents. *AAAI'2010*, 581-586.

28. Sun, B., Mitra, P., Giles, C., Yen, J. and Zha, H., 2007. Topic segmentation with shared topic detection and alignment of multiple documents. *SIGIR'07*, 199-206.

29. Tür, G., Stolcke, A., Hakkani-Tür, D. and Shriberg, E. 2001. Integrating prosodic and lexical cues for automatic topic segmentation, *Computational Linguistics,* 27(1), 31-57.

30. Wang, F., Tong, H. and Lin, C. 2011. Towards Evolutionary Nonnegative Matrix Factorization. *AAAI'11, 501-566.*

31. Wang, C., Blei, D. and Heckerman, D. 2008. Continuous Time Dynamic Topic Models. *Proc. on Uncertainty in AI*, 579-586.

32. Wang, F., Li, T. and Zhang, C. 2008. Semi-Supervised Clustering via Matrix Factorization. *SIAM'08*, 1-12.

33. Wang, X. and McCallum, A. 2006. Topics over time: a Non-Markov Continuous-Time Model of Topical Trends. *SIGKDD'06*, 424-433.