Evolutionary Timeline Summarization: a Balanced Optimization Framework via Iterative Substitution

Rui Yan
Dept. of Computer Science
Peking University
Beijing 100871, P. R. China
r.yan@pku.edu.cn

Xiaojun Wan
Institute of Computer Science and Technology
Peking University
Beijing 100871, P. R. China
wanxiaojun@icst.pku.edu.cn

Jahna Otterbacher
Lewis Department of Humanities
Illinois Institute of Technology
Chicago IL 60616, USA
jotterba@iit.edu

Liang Kong
Dept. of Machine Intelligence
Peking University
Beijing 100871, P. R. China
kongliang@pku.edu.cn

Xiaoming Li
Dept. of Computer Science
Peking University
Beijing 100871, P. R. China
lxm@pku.edu.cn

Yan Zhang*
Dept. of Machine Intelligence
Peking University
Beijing 100871, P. R. China
zhy@cis.pku.edu.cn

ABSTRACT

Classic news summarization plays an important role with the exponential document growth on the Web. Many approaches are proposed to generate summaries but seldom simultaneously consider evolutionary characteristics of news plus to traditional summary elements. Therefore, we present a novel framework for the web mining problem named Evolutionary Timeline Summarization (ETS). Given the massive collection of time-stamped web documents related to a general news query, ETS aims to return the evolution trajectory along the timeline, consisting of individual but correlated summaries of each date, emphasizing *relevance*, *coverage*, *coherence* and *cross-date diversity*. ETS greatly facilitates fast news browsing and knowledge comprehension and hence is a necessity. We formally formulate the task as an optimization problem via iterative substitution from a set of sentences to a subset of sentences that satisfies the above requirements, balancing coherence/diversity measurement and local/global summary quality. The optimized substitution is iteratively conducted by incorporating several constraints until convergence. We develop experimental systems to evaluate on 6 instinctively different datasets which amount to 10251 documents. Performance comparisons between different system-generated timelines and manually created ones by human editors demonstrate the effectiveness of our proposed framework in terms of ROUGE metrics.

*Corresponding author.

SIGIR'11, July 24–28, 2011, Beijing, China.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms

Algorithms, Experimentation, Performance

Keywords

Evolutionary summarization, timeline, optimization

1. INTRODUCTION

In the beginning, we will answer three possible questions.

Why timelines? The rapid growth of World Wide Web means that document floods spread throughout the Internet. Readers get drown in the sea of documents, wondering where to access. General search engines simply return webpages ranked by query relevance, but they are not quite capable of handling ambiguous intentioned queries, such as a query about evolving news "*Influenza A*". People may have the myriad of general interests about the beginning, the evolution or the most up to date situation, while IR techniques rank the returned webpages according to their understanding of relevance, which is insufficient.

In many cases, even if the ranked documents could be in a satisfying order, readers are tired of navigating every document in the overwhelming collection: they want to monitor the evolution trajectory of hot topics by simply browsing. Summarization is an ideal solution to provide a condensed, informative document reorganization for faster and better representation of news evolution. Timeline temporally summarizes evolutionary news as a series of *individual* but *correlated* component summaries and hence offers an option to understand the big picture of a developing situation.

Why not retrieve timelines created by editors? Manually generated timelines are concise, accurate and informative but require tremendous human labor of reading and the work is really energy consuming. Hence, it is impossible to generate timelines by hands for all queries from

Table 1: Part of human generated timeline about *Influenza A* outbreak in 2009 from Fox News website*.

February 2009
The first cases of an unusually aggressive cold are reported in the town of La Gloria in the Gulf coast state of Veracruz, Mexico.
April 2, 2009
The Veracruz government notifies authorities of a possible flu outbreak in the town of La Gloria. It is initially treated as a common respiratory illness.
April 25, 2009
The WHO announces that the world faces a swine flu pandemic.
April 27, 2009
The World Health Organization raises its pandemic alert status to Phase 4, meaning there is sustained human-to-human transmission of the virus causing outbreaks in at least one country.
April 29, 2009
The first U.S. death from swine flu is confirmed, a toddler from Mexico who traveled with family to Texas, where he died in the hospital on April 27. The WHO raises its pandemic alert status to Phase 5, meaning that a pandemic is imminent.
June 11, 2009
WHO declares the swine flu a pandemic, meaning that its spread is unstoppable.
July 21, 2009
The worldwide death toll passes 700, the WHO reports. An Australian pharmaceutical company says it will begin human trials of a swine flu vaccine.
Oct. 5, 2009
The first doses of swine flu vaccine, in nasal spray form, become available in the U.S. About 2.2 million doses are available.
Oct. 27, 2009
The CDC reports more than 22 million doses of swine flu vaccine are available.

*http://interactive.foxnews.com/health/swine-flu-timeline

users. Limited handcrafted timelines can be retrieved by search engines. Therefore, it is beneficial to automatically generate high-quality timelines from a collection of various news sources. The application can be embedded into search engines to improve users' retrieval experience. *Google News Timeline*[1] tries to provide such technique but it is coarse-grained, which merely clusters news articles into topic groups and sorts them chronologically [7]. Retrieved documents are neither summarized nor distinguished by their relevance, importance or informativeness.

Why not traditional summarization? A timeline is defined as a historical account of events ranged in chronological sequences. Unlike narratives, which select events in an interpretive context, timelines are more event-aware. Summarizing timelines is obviously different from traditional multi-document summarization (MDS). We first study a manual timeline of *Influenza A* epidemic spread trajectory in Table 1 from Fox News. We discover four prominent attributes from the timeline created by professional editors:

Relevance. As users issue queries for the news, they definitely need information relevant to what they query.

Coverage. To minimize the loss of main information, timelines should cover as many as possible important aspects during the news evolution, e.g. the first infected case, vaccine development, etc. illustrated in Table 1.

Coherence. Due to the dynamic nature of news over time, events on separate dates may share dependencies such as "follow-ups", "causes" or "consequences". The component summaries within the timeline capture such coherence to reflect the evolution trajectory, e.g. the escalated emergency, vaccine development from research to practical use, etc.

Diversity. Diversity has a double meaning: selected sentences tend to avoid information redundancy within a single date and a period of dates, namely cross-date diversity.

Unfortunately, traditional MDS neglects the significant temporal dimension for evolutionary summarization, nor does it incorporate news event characteristics into summarization, such as coherence, cross-date diversity and their balance by considering temporal proximity. Additionally, timeline consists several component summaries and the quality of these summaries should be optimized both locally, i.e., based on adjacent neighbors, and globally, i.e., based on the whole collection. Compared with traditional MDS without such concepts, timeline generation faces with new challenges.

We introduce a novel framework for the web mining service Evolutionary Timeline Summarization (ETS). Taking a general query issued by users and the returned collection as input, the system automatically outputs a timeline with items of component summaries which represent evolutionary trajectories on specific dates. According to the scores of timeline attributes, summaries are generated by ranking sentences in a balanced optimization framework through iterative substitution from a set of sentences to a subset of sentences under constraints. We build an experimental system on 6 real datasets to verify the effectiveness of our methods compared with 4 rivals. ETS addresses following challenges:

• The **1st challenge** for ETS is to model the the four attributes for component summaries: items are not assumed to be completely isolated because neighboring summaries are generated interdependently due to news characteristics over time. Diversity within an individual summary has been studied by others, but *cross-date diversity* has not previously been addressed. More importantly, *coherence* between component summaries has never been considered.

• We have global/local criteria to evaluate the qualities of component summaries. The **2nd challenge** is to formulate the task into a balanced optimization problem to generate summaries which satisfy double standards and above attributes. We propose an efficient framework via iterative substitution and enforce it through constraints construction.

• Due to ETS application scenario, the large-scale Web collection brings certain difficulties. Traditional MDS faces with a limited collection size (tens of documents) while ETS faces much larger corpora. A **3rd challenge** is to deal with significant corpus compression by filtering or pre-processing.

Our contributions are manifold by solving these challenges.

In Section 2 we start by reviewing previous works. In Section 3 we formulate ETS task as an optimization problem and define calculation functions based on four attributes. We explain the balanced optimization solution in Section 4 and describe the experiments in Section 5, including experimental system, performance comparisons, result discussion and case studies. Finally we draw conclusions in Section 6.

2. RELATED WORK

• Multi-document Summarization (MDS)

MDS has drawn much attention all these years and gained emphasis in workshops and conferences (SIGIR, ACL, DUC, etc.). General MDS can either be extractive or abstractive. The former assigns salient scores to semantic units (e.g. sentences, paragraphs) of the documents indicating their importance and then extracts top ranked ones, while the latter de-

[1]www.newstimeline.googlelabs.com

mands information fusion, such as sentence compression and reformulation. In this study we focus on extractive summarization to chronicle important news for timeline generation.

Centroid-based method is one of the most popular extractive summarization method. MEAD [14] and NeATS [11] are such implementations, using position, term frequency and theme, etc. MMR [6] algorithm is used to remove redundancy. Most recently, the graph-based ranking methods have been proposed to rank sentences or passages based on the "votes" or "recommendations" between each other. TextRank [13] and LexPageRank [3] use algorithms similar to PageRank and HITS to compute sentence importance. Wan *et al.* improve the graph-ranking algorithm by differentiating intra-document and inter-document links between sentences [17]. Cluster information has been incorporated in the graph model to better evaluate sentences [16]. Li *et al.* use a structural SVM to learn for sentence selection in MDS [9]. However, all these approaches are for traditional MDS and they miss the temporal dimension.

Swan and Allan construct timelines by extracting clusters of noun phrases and named entities [15]. Later they build a system to provide timelines which consist of one sentence per date, considering usefulness and novelty [1]. Chieu *et al.* build a similar system in units of sentences with interest and burstiness [2]. None of these methods enriches timeline measurement nor involves the evolutionary characteristics of news mentioned above, so we fill in the gaps by generating component summaries which are not completely independent: they have influence on "neighbors". ETS is based on a balanced optimization framework via iterative substitution.

• **Understanding News**

Topic detection and tracking (TDT) in news streams is extensively studied in the literature and identifies nature of news. Lexical similarity, temporal proximity and query relevance are introduced for topic detection as a part of task initiative, which is later combined with improved clustering techniques to establish event linkage. Novelty detection [10] is an important research branch to decide candidate event sentences. News correlation is determined by causal, temporal and rich semantic relationships in [4] or hierarchical dependencies in [5]. Features such as named entities, date or place information, and domain knowledge are deeply analyzed [21]. In this study, we do not seek to cluster "topics" like in TDT or in topic models but to utilize evolutionary correlations of news coherence/diversity for summarization.

3. PROBLEM FORMULATION

We give a formal definition of ETS as follows:

Input: Given a general query Q=$\{q_1, q_2, \ldots, q_{|Q|}\}$ from users where q_i is a query word, we obtain a sentence collection C from query related documents. We cluster the sentences into $\{C_1, C_2, \ldots, C_{|T|}\}$ by associated publish dates T=$\{t_1, t_2, \ldots, t_{|T|}\}$. t_i is the timestamp of sub-collection C_i.

Output: A evolutionary timeline which consists of a series of individual but correlated summary items, i.e. I=$\{I_1, I_2, \ldots, I_{|T|}\}$, where I_i on date t_i is a subset of C_i ($I_i \subseteq C_i$).

According to our investigation, we observe that an effective summary should properly consider the following four key requirements: (1)*Relevance.* Users are more interested in sentences which are related to the given query, which is similarly defined as "*interest*" in [2]; (2)*Coverage.* The summary should keep alignment with the source collection, which is proved to be significant as proposed in [9]; (3)*Coherence.* News changes gradually as time elapses and evolution indicates consistency among component summaries. It is a novel insight never considered before; (4)*Diversity.* According to MMR principle [6] and its applications [17, 16], a good summary should be concise and contain as few redundant sentences as possible, i.e., two sentences providing similar information should not be both present. Under our scenario, we extend to measure *cross-date diversity* as penalization to balance "coherence" and "diversity". All requirements involve a measurement of similarity between two word distributions Θ_1 and Θ_2, which are measured by Kullback-Leibler divergence here. We introduce decreasing/increasing logistic functions, $\mathcal{L}_1(x) = 1/(1+e^x)$ and $\mathcal{L}_2(x) = e^x/(1+e^x)$, to map the distance into interval [0,1]. V is the vocabulary set and $p(w|\Theta) = \frac{tf(w,\Theta)}{\sum_{w'} tf(w',\Theta)}$ where tf denotes the term frequency for word w.

$$D_{KL}(\Theta_1||\Theta_2) = \sum_{k \in V} p(w|\Theta_1) log \frac{p(w|\Theta_1)}{p(w|\Theta_2)}$$

Relevance. Given query Q, users are more interested in query-relevant information, namely *relevance* measured by $\mathcal{F}_r(I_i)$. However, Θ_Q is too sparse to reflect essential word distribution. Query expansion is introduced by pseudo-relevance feedback to enlarge Q. We retrieve top-κ snippets (semantic units of our system described in Section 5.2) relevant to Q to build a language model $\Theta_{Q'}$ as a worthy approximation to Θ_Q. Larger distance between Θ_{I_i} and $\Theta_{Q'}$ is not desired:

$$\mathcal{F}_r(I_i) = \mathcal{L}_1(D_{KL}(\Theta_{I_i}||\Theta_{Q'})). \quad (1)$$

Coverage. Summary I_i focuses on minimizing the loss of main information from sub-collection C_i. As ETS requires double criteria of summary quality, a good component summary should represent global source C as well to avoid probable local bias caused by discordant distributions of C_i and C. Coverage $\mathcal{F}_{cv}(I_i)$ is merged by an aggregation function for global coverage $\mathcal{F}_G(I_i) = \mathcal{L}_1(D_{KL}(\Theta_{I_i}||\Theta_C))$ and local coverage $\mathcal{F}_L(I_i) = \mathcal{L}_1(D_{KL}(\Theta_{I_i}||\Theta_{C_i}))$. We use a Jelinek-Mercer (JM) interpolation controlled by parameter λ:

$$\mathcal{F}_{cv}(I_i) = \lambda \cdot \mathcal{F}_L(I_i) + (1-\lambda) \cdot \mathcal{F}_G(I_i). \quad (2)$$

Coherence. As mentioned above, a timeline consists of a series of individual but correlated summaries. News evolves over time and a good component summary is coherent with neighboring summaries so that a timeline tracks the gradual evolution trajectory for multiple correlative news rather than the development by leaps and bounds. Therefore, we use $\mathcal{F}_{ch}(I_i)$ to evaluate *coherence* to measure the distance between Θ_{I_i} and the word distribution Θ_{N_i} from I_i's neighboring summary sets N_i.

$$\mathcal{F}_{ch}(I_i) = \mathcal{L}_1(D_{KL}(\Theta_{I_i}||\Theta_{N_i})). \quad (3)$$

Due to the scrutinized study of temporal proximity in news streams, terms on different dates are not equally weighted. Exponential decay is usually utilized to measure temporal distance [20]. Given $dtf(w, I_j|t_i) = e^{-\alpha|t_j - t_i|} \times tf(w, I_j)$, the decayed word distribution is calculated by:

$$p(w|\Theta_{N_i}) = \frac{\sum_I dtf(w, I_j|t_i)}{\sum_I \sum_{w' \in I_j}^{V} dtf(w', I_j|t_i)}.$$

Diversity. Traditional MDS shows a uniform tolerance towards redundancy to all candidate sentences, while in ETS diversity is dynamic: the tolerance arises as temporal gap

enlarges. Diversity measures the novelty degree of any of the sentence s compared with all other sentences not only within I_i, but also within other component summaries with decayed weights. Such *cross-date diversity* of an average novelty score $\mathcal{F}_d(I_i)$ is calculated by leaving out all sentences in I_i, one at a time. For diversity, larger distance is desired.

$$\mathcal{F}_d(I_i) = \frac{1}{|I_i|}\sum_{s\in I_i}\mathcal{L}_2(D_{KL}(\Theta_s||\Theta_{(N_i-s)})). \quad (4)$$

Utility. Given the source collection, the utility of an individual summary item I_i is evaluated based on the weighted combination of these requirements. All function values are between 0 and 1 and for simplicity, we let $\sum_k w_k = 1$.

$$\mathcal{U}(I_i) = w_1\mathcal{F}_r(I_i) + w_2\mathcal{F}_{cv}(I_i) + w_3\mathcal{F}_{ch}(I_i) + w_4\mathcal{F}_d(I_i) \quad (5)$$

Given the sentence set C_i and the compression rate ϕ_i on t_i, there are $\phi_i|C_i|$ out of $|C_i|$ possibilities to generate I_i. The ETS task is to predict the optimized sentence subset of I_i^* from the space of all combinations for all dates. The objective function is as follows:

$$I_i^* = \underset{I_i}{\operatorname{argmax}}\,\mathcal{U}(I_i). \quad (6)$$

As $\mathcal{U}(I_i)$ is measured by the neighboring summaries in the generated timeline in our framework, we generate I_i iteratively to approximate I_i^*, i.e, maximize $\mathcal{U}(I_i)$ based on the timeline generated in the last iteration.

4. OPTIMIZATION FRAMEWORK

4.1 Sentence Selection for Summaries

Based on the proposed metric of $\mathcal{U}(.)$, during each iteration the algorithm tends to highly score sentences which are more relevant to user interests, more aligned with source texts, more coherent with neighboring components generated in the last iteration and more diversified in the timeline. Hence top ranked sentences s according to $\mathcal{U}(s)$ are strong candidates for the target summaries.

Consider $I_i^{(n-1)}$ which consists of $\phi_i|C_i|$ sentences generated in the *(n-1)*-th iteration and the top $\phi_i|C_i|$ ranked sentences in the n-th iteration (denoted by $\mathcal{S}_i^{(n)}$), they have an intersection set of $\mathcal{Z}_i^{(n)} = I_i^{(n-1)} \cap \mathcal{S}_i^{(n)}$. There is a substitutable sentence set $\mathcal{X}_i^{(n)} = I_i^{(n-1)} - \mathcal{Z}_i^{(n)}$ and a new coming candidate sentence set $\mathcal{Y}_i^{(n)} = \mathcal{S}_i^{(n)} - \mathcal{Z}_i^{(n)}$. Under defined constraints, we substitute $\mathtt{x}_i^{(n)}$ sentences with $\mathtt{y}_i^{(n)}$, where $\mathtt{x}_i^{(n)} \subseteq \mathcal{X}_i^{(n)}$ and $\mathtt{y}_i^{(n)} \subseteq \mathcal{Y}_i^{(n)}$. During every iteration, our goal is to find a substitutive pair $< \mathtt{x}_i, \mathtt{y}_i >$ for I_i:

$$< \mathtt{x}_i, \mathtt{y}_i >: \mathcal{X}_i \times \mathcal{Y}_i \to \mathcal{R}_i.$$

To measure the performance of such substitution, a discriminant utility gain function

$$\begin{aligned}\Delta\mathcal{U}^{(n)}_{\mathtt{x}_i^{(n)},\mathtt{y}_i^{(n)}} &= \mathcal{U}(I_i^{(n)}) - \mathcal{U}(I_i^{(n-1)}) \\ &= \mathcal{U}((I_i^{(n-1)} - \mathtt{x}_i^{(n)}) \cup \mathtt{y}_i^{(n)}) - \mathcal{U}(I_i^{(n-1)})\end{aligned} \quad (7)$$

is employed to quantify the penalty. Therefore, we can predict the substitutive pair by maximizing the gain function $\Delta\mathcal{U}_{\mathtt{x}_i,\mathtt{y}_i}$ over the state set $\mathcal{R}_i$, with a size of

$$\sum_{k=0}^{\mathcal{Y}_i} A_{\mathcal{X}_i}^k C_{\mathcal{Y}_i}^k$$

where $<\mathtt{x}_i, \mathtt{y}_i> \in \mathcal{R}_i$. Finally the objective function of Equation (6) changes into maximization of utility gain by substitute $\hat{\mathtt{x}}_i$ with $\hat{\mathtt{y}}_i$ during each iteration. Formally,

$$< \hat{\mathtt{x}}_i, \hat{\mathtt{y}}_i > = \underset{\mathtt{x}_i\subseteq\mathcal{X}_i,\mathtt{y}_i\subseteq\mathcal{Y}_i}{\operatorname{argmax}} \Delta\mathcal{U}_{\mathtt{x}_i,\mathtt{y}_i}. \quad (8)$$

4.2 Balanced Optimization

A well generated summary I_i should be evaluated highly within the global collection C, and also be optimized given the local set of C_i. However, the objectives of two optimizations are not always the same because the word distributions in these source sets C and C_i are different. The substitutive pair <x, y> may perform well based on the timeline globally while not on the neighboring set locally and vice versa. There is a tradeoff between the global optimization and local optimization and hence we need to balance both.

Recall the aggregation function of linear combination between local and global coverage in Equation (2). The utility for I_i can also be rewritten as an interpolation function from local utility $\mathcal{U}(I_i)|_{C_i}$ from sub-collection C_i and global utility $\mathcal{U}(I_i)|_C$ from the global collection C:

$$\mathcal{U}(I_i) = \lambda \cdot \mathcal{U}(I_i)|_{C_i} + (1-\lambda)\cdot\mathcal{U}(I_i)|_C \quad (9)$$

The objective Equation (8) is actually to maximize $\Delta\mathcal{U}(I_i)$ from all possible substitutive pairs between two iterations to generate I_i. The algorithm is shown in Algorithm 1.

Algorithm 1 Interpolative Optimization

```
Input: C_1, C_2, ..., C_|T|, ϵ, φ_i
I_i ← {} for i=1, 2, ..., |T|
repeat
    for i = 1 to |T| do
        U'(I_i) = U(I_i)
        for all s ∈ C_i do
            calculate U(s)
        end for
        rank s with U(s)
        S_i ← top φ_i|C_i| ranked sentences
        Z_i ← I_i ∩ S_i
        X_i ← I_i − Z_i
        Y_i ← S_i − Z_i
        for all <x_i, y_i> pair where x_i ⊆ X_i, y_i ⊆ Y_i do
            ΔU_{x_i,y_i} = U((I_i − x_i) ∪ y_i) − U'(I_i)
        end for
        < x̂_i, ŷ_i >= argmax ΔU_{x_i,y_i}
        I_i ← (I_i − x̂_i) ∪ ŷ_i
        ΔU_i = U(I_i) − U'(I_i)
    end for
until ∀ ΔU_i < ϵ
```

The threshold ϵ is set at 0.0001 in this study. However, Algorithm 1 cannot avoid extreme situations. Significant rise in local utility which offsets much global utility loss still makes an available selection and vice versa. We seek to control such fluctuations to accelerate the convergence of the objective function. We choose the utility-maximized substitutive pairs under some constraints which are to ensure the overall utility is non-decreasing while finally both local and global optimizations are reached as iterations accumulate.

Local Optimization. The maximized substitutive pair should have utility improvement within sub-collection C_i. The local utility loss is not acceptable during the iterations.

Constraint 1.

$$\Delta\mathcal{U}_{\mathtt{x}_i,\mathtt{y}_i}|_{C_i} \geq 0 \tag{10}$$

Global Optimization. For each date, there can be a global utility loss for improvement in local utility, but with a borderline for such compromise. Global utility loss is acceptable for parts of the timeline but the loss from these dates should be offset by gains from the remaining dates. In other words, the sum of global utility is non-decreasing.

Constraint 2.

$$\sum_{I_i \in I} \Delta\mathcal{U}_{\mathtt{x}_i,\mathtt{y}_i}|_C \geq 0 \tag{11}$$

Constraint 3.

$$\Delta\mathcal{U}_{\mathtt{x}_i,\mathtt{y}_i}|_C > -\mu \qquad (\mu > 0) \tag{12}$$

μ is set as the maximum absolute value of $|(\Delta\mathcal{U}_{\mathtt{x}_i,\mathtt{y}_i}|_C)|$ from the last iteration and is initialized as positive infinity. Although allow utility compromises for parts of the timeline, we do not desire the situation of significant global utility gain for few dates along with slight utility loss for most others. Hence we set another constraint: for every iteration, the number of dates with global utility loss should not exceed m. In this study we let $m = \lfloor 0.8 \times |I| \rfloor$, which allows at most 80% dates with global utility loss.

Constraint 4.

$$|\{i|(\Delta\mathcal{U}_{\mathtt{x}_i,\mathtt{y}_i}|_C) < 0\}| < m \tag{13}$$

Optimization Problem. Considering the four types of constraints, we propose the balanced maximization framework enforcing both local and global optimization. Equation (8) can be rewritten as:

$$< \hat{\mathtt{x}}_i, \hat{\mathtt{y}}_i >= \underset{\mathtt{x}_i \subseteq \mathcal{X}_i, \mathtt{y}_i \subseteq \mathcal{Y}_i}{\operatorname{argmax}} \Delta\mathcal{U}_{\mathtt{x}_i,\mathtt{y}_i}, \tag{14}$$

subjected to:

$$(1)\Delta\mathcal{U}_{\mathtt{x}_i,\mathtt{y}_i}|_{C_i} \geq 0,$$
$$(2)\sum_{I_i \in I} \Delta\mathcal{U}_{\mathtt{x}_i,\mathtt{y}_i}|_C \geq 0,$$
$$(3)\Delta\mathcal{U}_{\mathtt{x}_i,\mathtt{y}_i}|_C > -\mu \quad (\mu > 0),$$
$$(4)|\{i|(\Delta\mathcal{U}_{\mathtt{x}_i,\mathtt{y}_i}|_C) < 0\}| < m.$$

Given results from Algorithm 1, suppose the size of the largest status space for a single date is $|H|$ within each iteration. We introduce a matrix $M_{|H|\times|T|}$ for all possible <x,y> pairs, where each element $M_{j,i}$ stands for a possible substitution. We calculate the global utility change $\Delta\mathcal{U}_{(M_{j,i})}|_C$, local utility change $\Delta\mathcal{U}_{(M_{j,i})}|_{C_i}$ and their linear combination $\Delta\mathcal{U}_{(M_{j,i})}$ for I_i. A **straightway understanding** is that we find a maximized overall utility $\Delta\mathcal{U}_{(M_{j,i})}$ at the j-th status space on date t_i, while at the same time global utility $\Delta\mathcal{U}_{(M_{j,i})}|_C$ and local utility $\Delta\mathcal{U}_{(M_{j,i})}|_{C_i}$ satisfy the four constraints. We select one element at each column by *Dynamic Programming*. After applying the substitution of $M_{j,i}$, a summary is hence generated within this iteration and the timeline is created by choosing a path in matrix $M_{|H|\times|T|}$.

We briefly describe the idea of dynamic programming and the details are left to Algorithm 2. Given matrix M, we sort all possible substitutions for I_i according to the overall utility change, i.e, $\Delta\mathcal{U}_{(M_{j,i})} > \Delta\mathcal{U}_{(M_{j+1,i})}$. We then split M into matrix M^L where elements $M^L_{j,i} = \Delta\mathcal{U}_{(M_{j,i})}|_{C_i}$ and matrix M^G where elements $M^G_{j,i} = \Delta\mathcal{U}_{(M_{j,i})}|_C$. In Algorithm 2 we set an array $\mathcal{A}[a][b][c]$= max $\{M_{j,a}\}$ where a is to record the processing column, b is to record how many $M^G_{j,i} < 0$ before column a on the path and c is to record the sum of $M^G_{j,i}$ before column a on the path. max $\{M_{j,a}\}$ denotes the maximum utility when specify a, b and c. Similar we set another array $\mathcal{P}[a][b][c]$ to record the path information. Details are illustrated in Algorithm 2. The worst time complexity is $O(|T| \times |H| \times m) \sim O(|T|^3)$ for Algorithm 2.

Algorithm 2 Dynamic programming with constraints

Input: Matrix M, M^L, M^G, m, μ
for $a = 0$ to $|T| - 1$ **do**
  for $b = 0$ to m **do**
    $c_{max} = max(\sum_{i=0}^{a} M^G_{j,i})$
    for $c = -\mu m$ to c_{max} **do**
      dynamic programming given $\mathcal{A}[a][b][c], \mathcal{P}[a][b][c]$
      for $l = 0$ to $|H|$ **do**
        if $M^L_{l,a+1} > 0$ && $0 > M^G_{l,a+1} > -\mu$ **then**
          sn=1
        else if $M^L_{l,a+1} > 0$ && $M^G_{l,a+1} > 0$ **then**
          sn=0
        end if
        if $\mathcal{A}[a+1][b+\text{sn}][c+M^G_{l,a+1}] < \mathcal{A}[a][b][c] + M_{l,a+1}$ **then**
          $\mathcal{A}[a+1][b+\text{sn}][c+M^G_{l,a+1}] = \mathcal{A}[a][b][c] + M_{l,a+1}$
          store path $\mathcal{P}[a+1][b+\text{sn}][c+M^G_{l,a+1}] = l_a$
        end if
      end for
    end for
  end for
end for
for $b = 0$ to m **do**
  $c_{max} = max(\sum_{i=0}^{|T|} M^G_{j,i})$
  for $c = 1$ to c_{max} **do**
    find maximum $\mathcal{A}[|T|][b][c]$
  end for
end for
trace way back by path $\mathcal{P}[|T|][b][c]$
return $l_{|T|}, \ldots, l_2, l_1$

5. EXPERIMENTS AND EVALUATION

5.1 Datasets

Since there is no existing standard test set for ETS methods, we construct 6 test sets which consist of news datasets and golden standards to evaluate our proposed framework empirically. We downloaded 10251 news articles from 10 selected sources. As shown in Table 2, one of the sources is in China, three of them are in UK and the rest are in the US. We choose them because many of these websites provide timelines edited by professional editors, which serve as golden standards. 6 topics belong to different categories of Rule of Interpretation (ROI) [8]. Statistics are in Table 3.

5.2 Experimental System Setups

We present 2 practical systems for ETS, *off-line* and *on-line*. Given a topic related corpus, the systems return trajectory timelines automatically. Off-line system handles stabilized topics with no new occurring while on-line system can support incremental documents from topics still evolving.

Table 2: News sources of 6 datasets

News Sources	Nation	News Sources	Nation
BBC	UK	Fox News	US
Xinhua	China	MSNBC	US
CNN	US	Guardian	UK
ABC	US	New York Times	US
Reuters	UK	Washington Post	US

Table 3: Detailed basic information of 6 datasets.

Topics (Query Words)	#Docs	#GT	AL	Since
1.Influenza A (H1N1)	2557	5	83	2009
2.Financial Crisis	2894	2	118	2009
3.BP Oil Spill	1468	6	76	2010
4.Haiti Earthquake	247	2	32	2010
5.Michael Jackson Death	925	3	64	2010
6.Obama Presidency	2160	5	92	2010

GT: ground truth; AL: average length of GT measured in sentences.

• **Preprocessing.** As ETS faces with much larger corpus compared with traditional MDS, we apply further data compression besides stemming and stop-word removal. We extract *text snippets* representing atomic "events" from all these documents with a toolkit provided by Yan *et al.* [19]. After the snippet extraction procedure, we compress the corpora by discarding non-event texts and filtering those events non-relevant to any of the query words.

• **Compression Rate.** After preprocessing, we obtain numerous snippets, temporally tagged according to the publish time of their source documents, and then decompose them into temporally tagged sentences as the global collection C. We partition C according to timestamps of sentences, i.e., $C = C_1 \cup C_2 \cup \cdots \cup C_{|T|}$. I_i is generated from sub-collection C_i. The sizes of component summaries are not necessarily equal. Users specify the overall compression rate ϕ, and we extract more sentences for important dates while fewer sentences for others. The *importance* of dates is measured by the *burstiness* with probable significant occurrences [2]. The compression rate on t_i is set as $\phi_i = \frac{|C_i|}{|C|}$.

• **Off-line System vs. On-line System.** The difference between two systems is whether corpora are temporally updating or not. For stabilized corpus, component summaries are optimized based on neighboring summaries on dates before and after them. For evolving corpus, we cannot forecast futures sentence sets, so the on-line system is to consider neighboring summaries previously generated.

5.3 Algorithms for Comparison

We implement the following widely used multi-document summarization algorithms as the baseline systems. Some of the systems are designed for traditional summarization without temporal dimension. The first intuitive generation for such methods is a global summarization on collection C at a uniform compression rate ϕ and then distribute the selected sentences to their source dates. The other intuitive one is a local summarization on sub-collection C_i with a compression rate ϕ_i. To these methods we take the average score as their performance. For fairness we conduct the same preprocessing for all algorithms by compression or filtering.

Random: The method selects sentences randomly for each document collection.

Centroid: The method applies MEAD algorithm [14] to extract sentences according to the following parameters: centroid value, positional value, and first-sentence overlap.

GMDS: The Graph-based MDS proposed by Wan *et al.* [16] first constructs a sentence connectivity graph based on cosine similarity and then selects important sentences based on the concept of eigenvector centrality.

Chieu: Chieu *et al.* present a similar timeline system [2] with different goals and frameworks, utilizing *interest* and *burstiness* ranking but neglecting news evolution.

ETS: Our proposed algorithms with iterative substitution under constraints are tested as ETS_1 for the off-line system and ETS_2 for the on-line system.

5.4 Evaluation Metrics

To compare with the human timelines, we use ROUGE toolkit (version 1.5.5), which is officially applied by Document Understanding Conference (DUC) for document summarization performance evaluation [12]. The summarization quality is measured by counting the number of overlapping units, such as N-gram, word sequences, and word pairs between the candidate timeline set CT and the ground-truth timelines GT. Several automatic evaluation methods are implemented in ROUGE, such as ROUGE-N, ROUGE-L and ROUGE-W, each of which can generate three scores (recall, precision and F-measure). Take ROUGE-N as an example:

1. ROUGE-N-R is an N-gram recall metric as follows:

$$\text{ROUGE-N-R} = \frac{\sum_{\text{I}\in\text{GT}} \sum_{\text{N-gram}\in\text{I}} \text{Count}_{\text{match}}(\text{N-gram})}{\sum_{\text{I}\in\text{GT}} \sum_{\text{N-gram}\in\text{I}} \text{Count (N-gram)}}$$

2. ROUGE-N-P is an N-gram precision metric as follows:

$$\text{ROUGE-N-P} = \frac{\sum_{\text{I}\in\text{CT}} \sum_{\text{N-gram}\in\text{I}} \text{Count}_{\text{match}}(\text{N-gram})}{\sum_{\text{I}\in\text{CT}} \sum_{\text{N-gram}\in\text{I}} \text{Count (N-gram)}}$$

3. ROUGE-N-F is an N-gram F_1 metric as follows:

$$\text{ROUGE-N-F} = \frac{2 \times \text{ROUGE-N-P} \times \text{ROUGE-N-R}}{\text{ROUGE-N-P} + \text{ROUGE-N-R}}$$

N in these metrics stands for the length of the N-gram and N-gram∈GT denotes the N-grams in the ground truth timeline while N-gram∈CT denotes the N-grams in the candidate timeline. $\text{Count}_{\text{match}}$(N-gram) is the maximum number of N-gram in the candidate summary and in the set of ground-truth summaries. $\text{Count}_{\text{(N-gram)}}$ is the number of N-grams in the ground truth summaries or candidate summary.

Furthermore, as the timeline consists of a series of individual summaries which are not equally significant, we evaluate ROUGE F-score for the timeline by the weighted average ROUGE F-score of all summaries, weighted by ϕ_i:

$$\text{ROUGE-N-F(I)} = \frac{1}{|\text{I}|} \frac{\sum_{I_i \in I} \phi_i \cdot \text{ROUGE-N-F}(\text{I}_\text{i})}{\sum_{I_i \in I} \sum_{I_k \in I} \phi_k \cdot \text{ROUGE-N-F}(\text{I}_\text{k})} \quad (15)$$

As we have similar conclusions in terms of any of the three scores, in this paper, we only report the average F-measure scores generated by unigram-based ROUGE-1, bigram-based ROUGE-2, and the weighted longest common subsequence based ROUGE-W to compare our proposed method with other implemented systems. These evaluation metrics have

been shown to much agree with human judgments. The weight W is set to be 1.2 in our experiments. Intuitively, the higher the ROUGE scores, the similar the two summaries.

5.5 Overall Performance Comparison

We use a cross validation manner among 6 datasets, i.e., we train parameters on one topic set and examine the performance on the others. After 6 training-testing processes, we take the average F-score performance in terms of ROUGE-1, ROUGE-2, and ROUGE-W on all sets. The overall results are shown in Figure 1 and details are listed in Tables 4∼6.

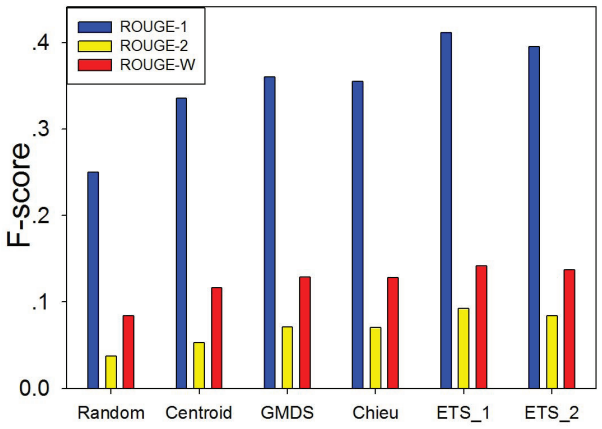


Figure 1: Overall performance on 6 datasets.

From the results, we have the following observations:

• Random has the worst performance as expected.

• The results of Centroid are better than those of Random. This is mainly because the Centroid based algorithm takes into account positional value and first-sentence overlap, which facilitates main aspects summarization.

• The GMDS system outperforms centroid-based summarization methods. This is due to the fact that PageRank-based framework ranks the sentence using eigenvector centrality which implicitly accounts for information subsumption among all sentences.

Traditional MDS only considers sentence selection from either the global or the local scope, and hence bias occurs. Many sentences are missed, which makes a low recall. Generally the performance of local priority summarization is better than global priority summarization. Probable bias is to some extent mitigated by searching for worthy sentence in every single date. However, precision drops due to excessive choice of local timeline-worthy sentences.

• In general, the result of Chieu's method is better than Centroid but unexpectedly, worse than GMDS. The reason in this case may be that Chieu's method does not capture sufficient timeline attributes. The "interest" modeled in their algorithms actually performs flat clustering-based summarization which is proved to be less useful [18]. GMDS utilizes sentence graph linkage, and partly captures "coherence".

• Both ETS_1 and ETS_2 under our proposed framework outperform baselines, indicating that the properties we use for timeline generation are beneficial. ETS_1 in off-line system performs better than ETS_2 in on-line system, indicating new coming documents do have influence on component summary generation within the timeline. ETS_2 is acceptable if on-line is required due to its advantage over baselines.

Table 4: Overall performance comparison on long lasting news. ROI* category: Science, Finance.

	1. Influenza A			2. Financial Crisis		
Systems	R-1	R-2	R-W	R-1	R-2	R-W
Random	0.257	0.039	0.081	0.230	0.030	0.071
Centroid	0.331	0.050	0.114	0.305	0.041	0.108
GMDS	0.364	0.062	0.130	0.327	0.054	0.110
Chieu	0.350	0.059	0.128	0.325	0.052	0.109
ETS_1	0.396	0.085	0.139	0.351	0.061	0.121
ETS_2	0.387	0.083	0.134	0.343	0.060	0.119

Table 5: Overall performance comparison on short breaking news. ROI category: Accidents, Disasters.

	3. BP Oil			4. Haiti Quake		
Systems	R-1	R-2	R-W	R-1	R-2	R-W
Random	0.262	0.041	0.096	0.266	0.043	0.093
Centroid	0.369	0.062	0.128	0.362	0.060	0.129
GMDS	0.389	0.084	0.139	0.380	0.106	0.137
Chieu	0.384	0.083	0.139	0.383	0.110	0.138
ETS_1	0.483	0.119	0.163	0.481	0.123	0.160
ETS_2	0.458	0.112	0.159	0.442	0.102	0.152

Table 6: Overall performance comparison on celebrities. ROI category: Legal Cases, Politics.

	5. Jackson Death			6. President Obama		
Systems	R-1	R-2	R-W	R-1	R-2	R-W
Random	0.232	0.033	0.080	0.254	0.039	0.084
Centroid	0.320	0.051	0.109	0.325	0.053	0.111
GMDS	0.341	0.059	0.127	0.359	0.061	0.129
Chieu	0.344	0.059	0.128	0.346	0.060	0.125
ETS_1	0.371	0.081	0.132	0.388	0.083	0.134
ETS_2	0.363	0.072	0.129	0.379	0.075	0.130

*ROI: news categorization defined by Linguistic Data Consortium. Available at http://www.ldc.upenn.edu/projects/tdt4/annotation

• The performance on intensive focused news within short time range (Topic 3, 4) is better than on long lasting news.

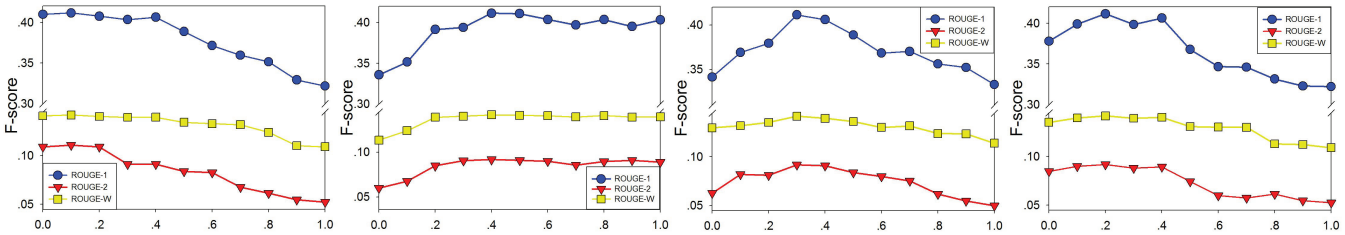
Having proved the effectiveness of our proposed methods, we carry the next move to identity how *relevance*, *coverage*, *coherence*, *diversity* and the 4 constraints take effects to enhance the quality of a summary in strategy selection.

5.6 Strategy Selection

Recall that utility $\mathcal{U}$ is the linear combination of local utility and global utility, both of which are the weighted sum of *relevance*, *coverage*, *coherence* and *diversity* under 4 constraints during the maximization process of $\mathcal{U}(I_i)$. Generally speaking, strategies can be sorted into two categories: parameter tuning and constraint selection. Each time, we tune one strategy while the other one is fixed.

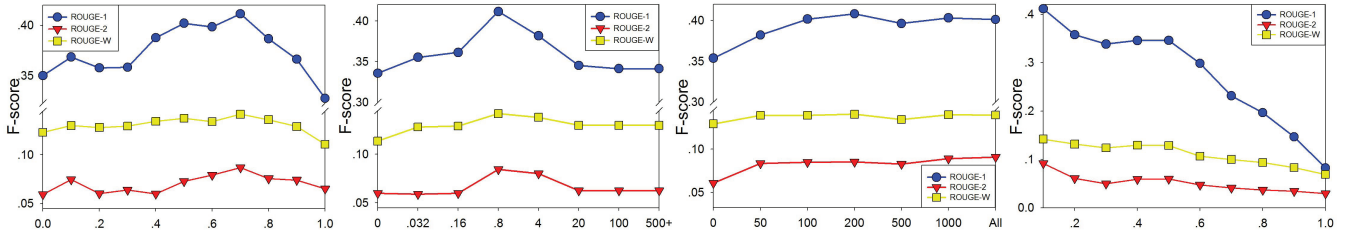
5.6.1 Parameter Tuning

Keeping other parameters fixed, we vary one parameter at a time to examine the changes of its performance from all 6 datasets. The first group of key parameters in our framework is w_1, w_2 and w_3 where w_4=1-w_1-w_2-w_3. Experimental results indicate coherence and diversity facilitate ETS while relevance demonstrates a relatively weaker influence. Excessive use of these 3 attributes impairs performance, except coverage, showing its domination in text summariza-



(a) w_1: effect of *relevance*. (b) w_2: effect of *coverage*. (c) w_3: effect of *coherence*. (d) w_4: effect of *diversity*.

Figure 2: Examine the performance of the four timeline-oriented attributes.



(e) λ: combine local/global balance. (f) α: control shape of decay. (g) κ: effect of pseudo-feedback. (h) ϕ: effect of compression rate.

Figure 3: Examine the performance of the experimental parameters.

tion. We set w_1=0.1, w_2=0.4, w_3=0.3 and hence w_4=0.2 in our experiments.

Another key parameter in our framework is λ in Equation (2) to measure the tradeoff between local and global collection. We gradually change λ from 0 to 1 at the step of 0.1 to examine the effect in Figure 3 (e). The combination of local and global utility outperforms the performance in isolation (λ=1 or 0). Furthermore, a larger λ (from 0.5 to 0.7) performs relatively better, but when λ exceeds 0.8, the extreme emphasis on global utility results in performance loss. We take λ=0.7 as the balance factor.

α controls the shape of the exponential decay and hence the size of influential neighboring window. We then examine the effect of neighboring summaries in ETS in Figure 3 (f). We vary α from 0 (all texts on timeline) to 500 (an approximation of $+\infty$, no neighbors considered). According to Figure 3 (f), the lines share a similar peak when $\alpha \in [0.8, 4]$. A moderate window size contributes to word distribution smoothing and reflects the trend for news evolution but too large a window introduces noise distribution as well. Therefore we choose α=0.8.

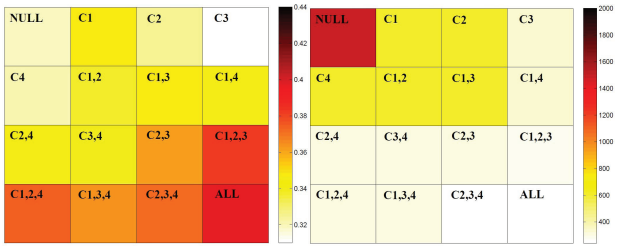
We then examine the results of different κ for pseudo-relevance feedback. According to Figure 3 (g), without any query expansion but simply compared with query Q, the performance is far from optimistic. Excessive document expansion impairs performance as well. κ=100 is shown large enough to smooth the word distribution in our experiments.

Finally we check the effect of overall compression rate ϕ which is usually designated by users. If the user would like to read more, he/she might favor a larger ϕ. We vary ϕ from 0.1 to 1 at the step of 0.1. Generally the lines are down-sloping as our ground-truth timelines are rather small compared with the huge global source collection. Recall is acceptable even when ϕ is small while precision drops accumulatively as ϕ increases.

5.6.2 *Constraints Selection*

To understand the effect of each proposed constraint, a series of experiments are conducted, illustrated in Figure 4, consisting all 2^4 combination tests of constraints C1∼C4.

From Figure 4, we notice Constraint 1 and Constraint 2 are useful. Recall the description of these two constraints. They are to maximize local utility gain and global utility gain and therefore they benefit timeline generation. The effectiveness of Constraint 3 and Constraint 4 seems not obvious in Figure 4 (a). Constraint 3 is to restrict the global utility loss for a particular summary. However, these two constraints do help reduce iteration counts to convergence, shown in Figure 4 (b). As iteration accumulates, the change of utility $\Delta\mathcal{U}$ varies significantly from time to time. It is difficult to set a general borderline of global utility loss arbitrarily to balance the convergence rate and timeline quality: inappropriate choice of Constraint 3 may cause potential harm to timeline generation. Both Constraint 3 and Constraint 4 are beneficial in iteration count performance because they reduce the available search space and facilitate early pruning for state paths in Algorithm 2.



(a). ROUGE-F-score. (b). Iteration counts.

Figure 4: ROUGE-F and convergence performance comparison among all 16 constraint combinations.

6. CONCLUSION

In this paper we present a novel framework for the important web mining problem named Evolutionary Timeline Summarization (ETS), which generates trajectory timelines

from massive data on the Internet. Given a query related news collection, ETS summarizes an evolution trajectory. We formally formulate ETS task as a balanced optimization problem via iterative substitution, measured on local sub-collections and the global collection. The objective function *Utility* is measured by four properties: *relevance*, *coverage*, *coherence* and *diversity*. We implement an off-/on-line system under such framework as experimental environment.

Abundant experiments are done on real web datasets. We compare numerous approaches, including two ways of implementation of 4 baselines and our proposed ETS_1, ETS_2. Through our experiments, we notice that among these properties, *coherence* plays an important role in timeline generation, indicating neighboring information is essential in evolutionary timeline trajectory: news evolves gradually. We also investigate the balance between local utility and global utility, and obtain the best combination coefficient at λ=0.7, meaning local utility weights slightly higher. We introduced four constraints, two of which ensure the local and global maximization, while the others ensure fast convergency. In case studies, our automatic timeline presents an informative document reorganization. However, as summaries generated by humans have potential biases, we will provide alternative evaluation metrics to measure ETS performance.

7. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable and constructive comments. This work is partially supported by NSFC Grant No.60933004, 61050009 and 61073081, and Xiaojun Wan is supported by NSFC Grant No.60873155.

8. REFERENCES

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 10–18, 2001.

[2] H. L. Chieu and Y. K. Lee. Query based event extraction along a timeline. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 425–432, 2004.

[3] G. Erkan and D. Radev. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of EMNLP*, volume 4, 2004.

[4] A. Feng and J. Allan. Finding and linking incidents in news. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 821–830, 2007.

[5] G. P. C. Fung, J. X. Yu, H. Liu, and P. S. Yu. Time-dependent event hierarchy construction. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 300–309, 2007.

[6] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: sentence selection and evaluation metrics. In *Proceedings of the 22nd SIGIR conference on Research and development in information retrieval*, pages 121–128, 1999.

[7] X. Jin, S. Spangler, R. Ma, and J. Han. Topic initiator detection on the world wide web. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 481–490, 2010.

[8] G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 297–304, 2004.

[9] L. Li, K. Zhou, G.-R. Xue, H. Zha, and Y. Yu. Enhancing diversity, coverage and balance for summarization through structure learning. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 71–80, 2009.

[10] X. Li and W. B. Croft. Improving novelty detection for general topics using sentence level information patterns. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 238–247, 2006.

[11] C.-Y. Lin and E. Hovy. From single to multi-document summarization: a prototype system and its evaluation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 457–464, 2002.

[12] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of NAACL'03*, pages 71–78, 2003.

[13] R. Mihalcea and P. Tarau. A language independent algorithm for single and multiple document summarization. In *Proceedings of IJCNLP*, 2005.

[14] D. Radev, H. Jing, M. Sty, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938, 2004.

[15] R. Swan and J. Allan. Automatic generation of overview timelines. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 49–56, 2000.

[16] X. Wan and J. Yang. Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 299–306, 2008.

[17] X. Wan, J. Yang, and J. Xiao. Single document summarization with document expansion. In *AAAI*, pages 931–936, 2007.

[18] D. Wang and T. Li. Document update summarization using incremental hierarchical clustering. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 279–288, 2010.

[19] R. Yan, Y. Li, Y. Zhang, and X. Li. Event recognition from news webpages through latent ingredients extraction. *Information Retrieval Technology*, pages 490–501, 2010.

[20] C. C. Yang and X. Shi. Discovering event evolution graphs from newswires. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 945–946, 2006.

[21] K. Zhang, J. Zi, and L. G. Wu. New event detection based on indexing-tree and named entity. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 215–222, 2007.

Table 7: Selected part of timeline generated by balanced ETS with optimization for $H1N1$.

April 23, 2009 s_1: The WHO makes its first report about the so-called Swine Influenza (A) H1N1. s_2: In Mexico, more than a thousand people have been infected; WHO has sent a team of experts there to further study the outbreak.	**April 29, 2009** s_1: President Obama asks Congress for $1.5 billion to fight the swine flu outbreak, build drug stockpiles and monitor future cases. s_2: WHO's decision to raise the alert level helps mobilize pharmaceutical companies and governments to start manufacturing adequate antiviral drugs and speed up the creation of a vaccine. s_3: For the first time, the World Health Organization raises the pandemic level to 5. s_4: Egypt begins slaughtering the country's roughly 300,000 pigs as a precaution.
April 25, 2009 s_1: WHO holds emergency meeting, says outbreak has potential to become a pandemic. s_2: The Mexican Director-General declares Swine Influenza (A) H1N1 a public health emergency. s_3: U.S. emergency departments step up efforts to control the virus should it surface. s_4: More than 100 students are sick with flulike symptoms.	**October 2, 2009** s_1: The US announces implementation of a massive campaign to vaccinate millions of Americans against swine flu, with the first 600,000 doses to be distributed in coming days. s_2: The US ordered 979 million dollars worth of Fluviron H1N1 vaccine from Novaritis. s_3: Spray vaccines will be the first to reach vaccination sites, one of the most vulnerable groups, pregnant women, will have to wait until later this month for the injection version.
April 27, 2009 s_1: The World Health Organization raises the pandemic alert one level to phase 4, which is two steps short of declaring a full-blown pandemic. s_2: A general practitioner considered swine flu a possible diagnosis, but specimens were not stored properly and a laboratory assessment could not confirm the case. s_3: European Union's health commissioner warns Europeans to avoid nonessential travel to Mexico and the United States. s_4: The World Bank in Washington, D.C., says a staff member who traveled to Mexico on business April 14-18 has been "preliminarily diagnosed" with swine flu. s_5: 74 schools are closed, leaving students out of classes by swine flu.	**October 5, 2009** s_1: UN officials warn that poor countries face "explosive outbreaks" of the global swine flu pandemic and need speedy financial assistance to access vaccines. s_2: WHO said pharmaceutical firms can produce only 3 billion doses of H1N1 vaccines a year, covering less than half of the global population.

Table 8: Selected part of timeline generated by balanced ETS with optimization for $BP\ Oil$.

April 20, 2010 s_1: Explosion and fire on the BP-licensed Transocean drilling rig Deepwater Horizon in the Gulf of Mexico. s_2: Deepwater Horizon oil rig fire leaves 11 missing. s_3: The rig was drilling in about 5,000ft (1,525m) of water, pushing the boundaries of deepwater drilling technology. s_4: A blowout preventer, intended to prevent release of crude oil, failed to activate.	**April 24, 2010** s_1: Oil is found to be leaking from the well. **April 26, 2010** s_1: BP's shares fall 2% amid fears that the cost of cleanup and legal claims will hit the London-based company hard. s_2: Roughly 15,000 gallons of dispersants and 21,000ft of containment boom are placed at the spill site.
April 22, 2010 s_1: The Deepwater Horizon sinks to the bottom of the Gulf after burning for 36 hours, raising concerns of a catastrophic oil spill. s_2: Deepwater Horizon rig sinks in 5,000ft of water. s_3: Reports of a five-mile-long oil slick. Search-and-rescue operations by the US National Response Team begin.	**April 27, 2010** s_1: The US departments of interior and homeland security announce plans for a joint investigation of the explosion and fire. s_2: Oil spill to be set on fire to save US coast. s_3: Minerals Management Service (MMS) approves a plan for two relief wells. s_4: BP reports a rise in profits, due in large part to oil price increases, as shares rise again.
April 23, 2010 s_1: The US coast guard suspends the search for missing workers, who are all presumed dead. s_2: The rig is found upside down about a quarter-mile from the blowout preventer. s_3: The Coast Guard says it had no indication that oil was leaking from the well 5,000ft below the surface of the Gulf. s_4: Underwater robots try to shut valves on the blowout preventer to stop the leak, but BP abandons that failed effort two weeks later. s_5: Deepwater Horizon clean-up workers fight to prevent disaster.	**April 28, 2010** s_1: The US Coast Guard warns the oil leak could become the worst oil spill in US history. s_2: The coast guard says the flow of oil is 5,000bpd, five times greater than first estimated, after a third leak is discovered. s_3: Controlled burns begin on the giant oil slick. s_4: BP's attempts to repair a hydraulic leak on the blowout preventer valve are unsuccessful.

Table 9: Selected part of timeline generated by balanced ETS with optimization for $Obama$.

January 15, 2010 s_1: US President Barack Obama spoke by telephone with Haitian President Rene Preval Friday morning, pledging full support of the United States in the ongoing earthquake relief effort. s_2: US takes charge in Haiti_with troops, rescue aid.	**June 26, 2010** s_1: The President holds separate bilateral meetings with Prime Minister David Cameron of the United Kingdom, President Lee Myung-bak of the Republic of Korea, President Hu Jintao of the People's Republic of China and attends the G20 Working Dinner.
March 19, 2010 s_1: US President Barack Obama makes remarks on health care reform at George Mason University in Fairfax, Virginia.	**November 2, 2010** s_1: WASHINGTON–President Obama holds a news conference in the White House to acknowledge that he and the Democratic party took a "shellacking" in the mid-term elections. s_2: Republicans rolled to their greatest midterms gains in 80 years, recapturing the House of Representatives and cutting the Democrats' majority in the Senate. s_3: Midterm election results show voters unhappy with President Obama's leadership.
March 21, 2010 s_1: Obama's presidency hinges on historic health care reform vote; If it passes, it's salvaged. s_2: The United States House of Representatives will vote on President Barack Obama's healthcare reform bill, and either way, shockwaves will be heard throughout the globe.	
March 23, 2010 s_1: Vice President Joe Biden introduces President Barack Obama on March 23, before the president signed the health care reform bill. s_2: Doubted and deeply in need of a comeback, President Barack Obama had a political dream week: a historic remaking of America's health care system, an overhaul of how students pay for college and a groundbreaking deal with Russia to shrink nuclear arsenals. s_3: Speaker Nancy Pelosi released the following statement today after President Barack Obama signed the Senate health insurance reform legislation into law.	**November 12, 2010** s_1: The president and the first lady are in the midst of a 10-day visit to Asia, the longest foreign trip of the Obama presidency thus far. s2: After what Mr. Obama termed a "shellacking," he pronounced himself ready to cooperate with Republicans. **December 3, 2010** s_1: (Reuters)-President Barack Obama makes a trip to Afghanistan to visit with US troops, and bad weather forced him to cancel a planned face-to-face meeting with Afghan President Hamid Karzai. s_2: President Obama grants his first presidential pardons to 9 people.