Timeline Generation through Evolutionary Trans-Temporal Summarization

Rui Yan[†], Liang Kong[†] , Congrui Huang[†], Xiaojun Wan[‡], Xiaoming Li[♮], Yan Zhang[†*]
[†]School of Electronics Engineering and Computer Science, Peking University, China
[‡]Institute of Computer Science and Technology, Peking University, China
[♮]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China
{r.yan,kongliang,hcr,lxm}@pku.edu.cn,
wanxiaojun@icst.pku.edu.cn,zhy@cis.pku.edu.cn

Abstract

We investigate an important and challenging problem in summary generation, i.e., Evolutionary Trans-Temporal Summarization (ETTS), which generates news timelines from massive data on the Internet. ETTS greatly facilitates fast news browsing and knowledge comprehension, and hence is a necessity. Given the collection of time-stamped web documents related to the evolving news, ETTS aims to return news evolution along the timeline, consisting of individual but correlated summaries on each date. Existing summarization algorithms fail to utilize trans-temporal characteristics among these component summaries. We propose to model trans-temporal correlations among component summaries for timelines, using *inter-date* and *intra-date* sentence dependencies, and present a novel combination. We develop experimental systems to compare 5 rival algorithms on 6 instinctively different datasets which amount to 10251 documents. Evaluation results in ROUGE metrics indicate the effectiveness of the proposed approach based on trans-temporal information.

1 Introduction

Along with the rapid growth of the World Wide Web, document floods spread throughout the Internet. Given a large document collection related to a news subject (for example, *BP Oil Spill*), readers get lost in the sea of articles, feeling confused and powerless. General search engines can rank these news webpages by *relevance* to a user specified aspect, i.e., a query such as "*first relief effort for BP Oil Spill*", but search engines are not quite capable of ranking documents given the whole news subject without particular aspects. Faced with thousands of news documents, people usually have a myriad of interest aspects about the beginning, the development or the latest situation. However, traditional information retrieval techniques can only rank webpages according to their understanding of relevance, which is obviously insufficient (Jin et al., 2010).

Even if the ranked documents could be in a satisfying order to help users understand news evolution, readers prefer to monitor the evolutionary trajectories by simply browsing rather than navigate every document in the overwhelming collection. Summarization is an ideal solution to provide an abbreviated, informative reorganization for faster and better representation of news documents. Particularly, a timeline (see Table 1) can summarize evolutionary news as a series of *individual* but *correlated* component summaries (items in Table 1) and offer an option to understand the big picture of evolution.

With unique characteristics, summarizing timelines is significantly different from traditional summarization methods which are awkward in such scenarios. We first study a manual timeline of BP Oil Spill in Mexico Gulf in Table 1 from Reuters News[1] to understand why timelines generation is observably different from traditional summarization. No traditional method has considered to partition corpus into subsets by timestamps for trans-temporal correlations. However, we discover two unique trans-

*Corresponding author.

[1]http://www.reuters.com

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 433–443,
Edinburgh, Scotland, UK, July 27–31, 2011.

Table 1: Part of human generated timeline about BP Oil Spill in 2010 from Reuters News website.

April 22, 2010
The Deepwater Horizon rig, valued at more than $560 million, sinks and a five mile long (8 km) oil slick is seen.
April 25, 2010
The Coast Guard approves a plan to have remote underwater vehicles activate a blowout preventer and stop leak. Efforts to activate the blowout preventer fail.
April 28, 2010
The Coast Guard says the flow of oil is 5,000 barrels per day (bpd) (210,000 gallons/795,000 litres) – five times greater than first estimated. A controlled burn is held on the giant oil slick.
April 29, 2010
U.S. President Barack Obama pledges "every single available resource," including the U.S. military, to contain the spreading spill. Obama also says BP is responsible for the cleanup. Louisiana declares state of emergency due to the threat to the state's natural resources.
April 30, 2010
An Obama aide says no drilling will be allowed in new areas, as the president had recently proposed, until the cause of the Deepwater Horizon accident is known.

temporal characteristics of component summaries from the handcrafted timeline. **Individuality.** The component summaries are summarized *locally*: the component item on date t is constituted by sentences with timestamp t. **Correlativeness.** The component summaries are correlative across dates, based on the *global* collection. To the best of our knowledge, no traditional method has examined the relationships among these timeline items.

Although it is profitable, summarizing timeline faces with new challenges:

• The first challenge for timeline generation is to deliver important contents and avoid information overlaps among component summaries under the trans-temporal scenario based on global/local source collection. Component items are individual but not completely isolated due to the dynamic evolution.

• As we have *individuality* and *correlativeness* to evaluate the qualities of component summaries, both locally and globally, the second challenge is to formulate the combination task into a balanced optimization problem to generate the timelines which satisfy both standards with maximum utilities.

We introduce a novel approach for the web mining problem Evolutionary Trans-Temporal Summarization (ETTS). Taking a collection relevant to a news subject as input, the system automatically outputs a timeline with items of component summaries which represent evolutionary trajectories on specific dates. We classify sentence relationships as *inter-date* and *intra-date* dependencies. Particularly, the inter-date dependency calculation includes temporal decays to project sentences from all dates onto the same time horizon (Figure 1 (a)). Based on intra-/inter-date sentence dependencies, we then model affinity and diversity to compute the saliency score of each sentence and merge local and global rankings into one unified ranking framework. Finally we select top ranked sentences. We build an experimental system on 6 real datasets to verify the effectiveness of our methods compared with other 4 rivals.

2 Related Work

Multi-document summarization (MDS) aims to produce a summary delivering the majority of information content from a set of documents and has drawn much attention in recent years. Conferences such as ACL, SIGIR, EMNLP, etc., have advanced the technology and produced several experimental systems.

Generally speaking, MDS methods can be either extractive or abstractive summarization. Abstractive summarization (e.g. NewsBlaster[2]) usually needs information fusion, sentence compression and reformulation. We focus on extraction-based methods, which usually involve assigning saliency scores to some units (e.g. sentences, paragraphs) of the documents and extracting the units with highest scores.

To date, various extraction-based methods have been proposed for generic multi-document summarization. The centroid-based method MEAD (Radev et al., 2004) is an implementation of the centroid-based method that scores sentences based on features such as cluster centroids, position, and TF.IDF, etc. NeATS (Lin and Hovy, 2002) adds new features such as topic signature and term clustering to select important content, and use MMR (Goldstein et al., 1999) to remove redundancy.

Graph-based ranking methods have been proposed to rank sentences/passages based on "votes" or "recommendations" between each other. TextRank (Mihalcea and Tarau, 2005) and LexPageRank (Erkan and Radev, 2004) use algorithms similar to PageRank and HITS to compute sentence importance. Wan et al. have improved the graph-ranking

[2]http://www1.cs.columbia.edu/nlp/newsblaster/

algorithm by differentiating intra-document and inter-document links between sentences (2007b), and have proposed a manifold-ranking method to utilize sentence-to-sentence and sentence-to-topic relationships (Wan et al., 2007a).

ETTS seems to be related to a very recent task of "update summarization" started in DUC 2007 and continuing with TAC. However, update summarization only dealt with a single update and we make a novel contribution with multi-step evolutionary updates. Further related work includes similar timeline systems proposed by (Swan and Allan, 2000) using named entities, by (Allan et al., 2001) measured in *usefulness* and *novelty*, and by (Chieu and Lee, 2004) measured in *interest* and *burstiness*. We have proposed a timeline algorithm named "Evolutionary Timeline Summarization (ETS)" in (Yan et al., 2011b) but the refining process based on generated component summaries is time consuming. We aim to seek for more efficient summarizing approach.

To the best of our knowledge, neither update summarization nor traditional systems have considered the relationship among "component summaries", or have utilized trans-temporal properties. ETTS approach can also naturally and simultaneously take into account global/local summarization with biased information richness and information novelty, and combine both summarization in optimization.

3 Trans-temporal Summarization

We conduct trans-temporal summarization based on the global biased graph using *inter-date* dependency and local biased graph using *intra-date* dependency. Each graph is the complementary graph to the other.

3.1 Global Biased Summarization

The intuition for global biased summarization is that the selected summary should be correlative with sentences from neighboring dates, especially with those informative ones. To generate the component summary on date t, we project all sentences in the collection onto the time horizon of t to construct a global affinity graph, using temporal decaying kernels.

3.1.1 Temporal Proximity Based Projection

Clearly, a major technical challenge in ETTS is how to define the temporal biased projection function $\Gamma(\Delta t)$, where Δt is the distance between the pending date t and neighboring date t', i.e., $\Delta t = |t' - t|$. As in (Lv and Zhai, 2009), we present 5 representative kernel functions: Gaussian, Triangle, Cosine, Circle, and Window, shown in Figure 2. Different kernels lead to different projections.

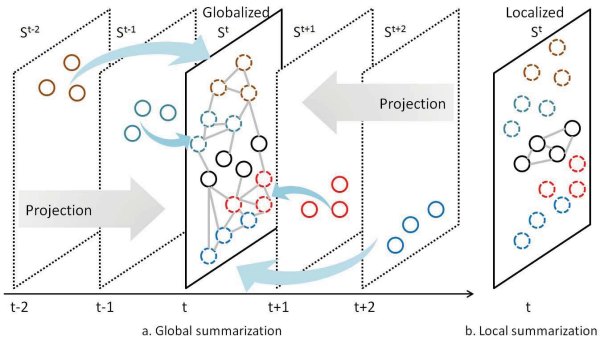


Figure 1: Construct global/local biased graphs. Solid circles denote intra-date sentences on the pending date t and dash ones represent inter-date sentences from other dates.

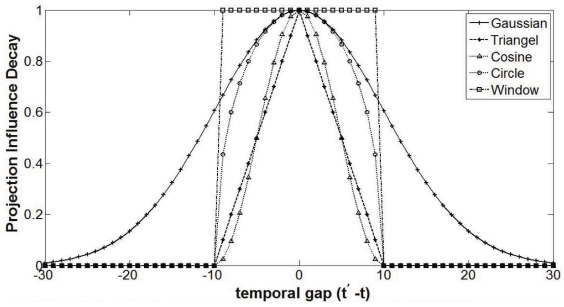


Figure 2: Proximity-based kernel functions, where σ=10.

1. Gaussian kernel

$$\Gamma(\Delta t) = \exp[\frac{-\Delta t^2}{2\sigma^2}]$$

2. Triangle kernel

$$\Gamma(\Delta t) = \begin{cases} 1 - \frac{\Delta t}{\sigma} & \text{if } \Delta t \leq \sigma \\ 0 & \text{otherwise} \end{cases}$$

3. Cosine (Hamming) kernel

$$\Gamma(\Delta t) = \begin{cases} \frac{1}{2}[1 + \cos(\frac{\Delta t \cdot \pi}{\sigma})] & \text{if } \Delta t \leq \sigma \\ 0 & \text{otherwise} \end{cases}$$

4. Circle kernel

$$\Gamma(\Delta t) = \begin{cases} \sqrt{1 - (\frac{\Delta t}{\sigma})^2} & \text{if } \Delta t \leq \sigma \\ 0 & \text{otherwise} \end{cases}$$

5. Window kernel

$$\Gamma(\Delta t) = \begin{cases} 1 & \text{if } \Delta t \leq \sigma \\ 0 & \text{otherwise} \end{cases}$$

All kernels have one parameter σ to tune, which controls the spread of kernel curves, i.e., it restricts the projection scope of each sentence. In general, the optimal setting of σ may vary according to the news set because sentences presumably would have wider semantic scope in certain news subjects, thus requiring a higher value of σ and vice versa.

3.1.2 Modeling Global Affinity

Given the sentence collection C partitioned by the timestamp set T, $C = \{C^1, C^2, \ldots, C^{|T|}\}$, we obtain $C^t = \{s_i^t | 1 \leq i \leq |C^t|\}$ where s_i is a sentence with the timestamp $t = t_{s_i}$. When we generate component summary on t, we project all sentences onto time horizon t. After projection, all sentences are weighted by their influence on t. We use an affinity matrix M^t with the entry of the inter-date transition probability on date t. The sum of each row equals to 1. Note that for the global biased matrix, we measure the affinity between local sentences from t and global sentences from other dates. Therefore, intra-date transition probability between sentences with the timestamp t is set to 0 for local summarization.

$M_{i,j}^t$ is the transition probability of s_i to s_j based on the perspective of date t, i.e., $p(s_i \rightarrow s_j|t)$:

$$p(s_i \rightarrow s_j|t) = \begin{cases} \frac{f(s_i \rightarrow s_j|t)}{\sum_{|C|} f(s_i \rightarrow s_k|t)} & \text{if } \sum f \neq 0 \\ 0 & \text{if } t_{s_i} = t_{s_j} = t \end{cases} \quad (1)$$

$f(s_i \rightarrow s_j|t)$ is defined as the temporal weighted cosine similarity between two sentences:

$$f(s_i \rightarrow s_j|t) = \sum_{w \in s_i \cap s_j} \pi(w, s_i|t) \cdot \pi(w, s_j|t) \quad (2)$$

where the weight π associated with term w is calculated with the temporal weighted *tf.isf* formula:

$$\pi(w, s|t) = \frac{\Gamma|t - t_s| \cdot tf(w, s)(1 + log(\frac{|C|}{N_w}))}{\sqrt{\sum_{|s|}(tf(w, s)(1 + log(\frac{|C|}{N_w})))^2}}. \quad (3)$$

where t_s is the timestamp of sentence s, and $tf(w, s)$ is the term frequency of w in s. t_s can be any date from T. $|C|$ is the sentences set size and N_w is the number of sentences containing term w.

We let $p(s_i \rightarrow s_i|t)$=0 to avoid self transition. Note that although $f(.)$ is a symmetric function, $p(s_i \rightarrow s_j|t)$ is usually not equal to $p(s_j \rightarrow s_i|t)$, depending on the degrees of nodes s_i and s_j.

Now we establish the affinity matrix $M_{i,j}^t$ and by using the general form of PageRank, we obtain:

$$\vec{\lambda} = \mu M^{-1} \vec{\lambda} + \frac{1 - \mu}{|C|} \vec{e} \quad (4)$$

where $\vec{\lambda}$ is the selective probability of all sentence nodes and $\vec{e}$ is a column vector with all elements equaling to 1. μ is the damping factor set as 0.85. Usually the convergence of the iteration algorithm is achieved when difference between the scores computed at two successive iterations for any sentences falls below a given threshold (0.0001 in this study).

3.1.3 Modeling Diversity

Diversity is to reflect both biased information richness and sentence novelty, which aims to reduce information redundancy. However, using standard PageRank of Equation (4) will not result in diversity. The aggregational effect of PageRank assigns high salient scores to closely connected node communities (Figure 3 (b)). A greedy vertex selection algorithm may achieve diversity by iteratively selecting the most prestigious vertex and then penalizing the vertices "covered" by the already selected ones, such as Maximum Marginal Relevance and its applications in Wan et al. (2007b; 2007a). Most recently diversity rank $DivRank$ is another solution to diversity penalization in (Mei et al., 2010).

We incorporate DivRank in our general ranking framework, which creates a dynamic M during each iteration, rather than a static one. After z times of iteration, the matrix M becomes:

$$M^{(z)} = \mu M^{(z-1)} \cdot \vec{\lambda}^{(z-1)} + \frac{1 - \mu}{|C|} \vec{e} \quad (5)$$

Equation (5) raises the probability for nodes with higher centrality and nodes already having high weights are likely to "absorb" the weights of its neighbors directly, and the weights of neighbors' neighbors indirectly. The process is to iteratively adjust matrix M according to $\vec{\lambda}$ and then to update $\vec{\lambda}$ according to the changed M. As iteration increases

there emerges a **rich-gets-richer** phenomenon (Figure 3 (c) and (d)). By incorporating DivRank, we obtain rank $r_i^\dagger$ and the global biased ranking score $\mathcal{G}_i$ for sentence s_i from date t to summarize C^t.

3.2 Local Biased Summarization

Naturally, the component summary for date t should be informative within C^t. Given the sentence collection $C^t = \{s_i^t | 1 \leq i \leq |C^t|\}$, we build an affinity matrix for Figure 1 (b), with the entry of intra-date transition probability calculated from standard cosine similarity. We incorporate DivRank within local summarization and we obtain the local biased rank and ranking score for s_i, denoted as $r_i^\ddagger$ and $\mathcal{L}_i$.

3.3 Optimization of Global/Local Combination

We do not directly add the global biased ranking score and local biased ranking score, as many previous works did (Wan et al., 2007b; Wan et al., 2007a), because even the same ranking score gap may indicate different rank gaps in two ranking lists.

Given subset C^t, let $R = \{r_i\}$(i = 1,...,$|C^t|$), r_i is the final ranking of s_i to estimate, optimize the following objective cost function $O(R)$,

$$\begin{aligned} O(R) =& \alpha \sum_{i=1}^{|C^t|} \mathcal{G}_i \| \frac{r_i}{\Psi_i} - \frac{r_i^\dagger}{\mathcal{G}_i} \|^2 \\ &+ \beta \sum_{i=1}^{|C^t|} \mathcal{L}_i \| \frac{r_i}{\Psi_i} - \frac{r_i^\ddagger}{\mathcal{L}_i} \|^2 \end{aligned} \tag{6}$$

where $\mathcal{G}_i$ is the global biased ranking score while $\mathcal{L}_i$ is the local biased ranking score. Ψ_i is expected to be the merged ranking score, namely sentence *importance*, which will be defined later. Among the two components in the objective function, the first component means that the refined rank should not deviate too much from the global biased rank. We use $\| \frac{r_i}{\Psi_i} - \frac{r_i^\dagger}{\mathcal{G}_i} \|^2$ instead of $\|r_i - r_i^\dagger\|^2$ in order to distinguish the differences between sentences from the same rank gap. The second component is similar by refining rank from local biased summarization.

Our goal is to find $R = R^*$ to minimize the cost function, i.e., $R^* = argmin\{O(R)\}$. R^* is the final rank merged by our algorithm. To minimize $O(R)$, we compute its first-order partial derivatives.

$$\frac{\partial O(R)}{\partial r_i} = \frac{2\alpha}{\Psi_i}\left(\frac{\mathcal{G}_i}{\Psi_i} r_i - r_i^\dagger\right) + \frac{2\beta}{\Psi_i}\left(\frac{\mathcal{L}_i}{\Psi_i} r_i - r_i^\ddagger\right) \tag{7}$$

Let $\frac{\partial O(R)}{\partial r_i} = 0$, we get

$$r_i^* = \frac{\alpha \Psi_i r_i^\dagger + \beta \Psi_i r_i^\ddagger}{\alpha \mathcal{G}_i + \beta \mathcal{L}_i} \tag{8}$$

Two special cases are that if (1) $\alpha = 0$, $\beta \neq 0$: we obtain $r_i = \Psi_i r_i^\ddagger / \mathcal{L}_i$, indicating we only use the local ranking score. (2) $\alpha \neq 0$, $\beta = 0$, indicating we ignore local ranking score and only consider global biased summarization using inter-date dependency.

There can be many ways to calculate the sentence importance Ψ_i. Here we define Ψ_i as the weighted combination of itself with ranking scores from global biased and local biased summarization:

$$\Psi_i^{(z)} = \frac{\alpha \mathcal{G}_i + \beta \mathcal{L}_i + \gamma \Psi_i^{(z-1)}}{\alpha + \beta + \gamma}. \tag{9}$$

To save one parameter we let $\alpha + \beta + \gamma = 1$. In the z-th iteration, $r_i^{(z)}$ is dependent on $\Psi_i^{(z-1)}$ and $\Psi_i^{(z)}$ is indirectly dependent on $r_i^{(z)}$ via $\Psi_i^{(z-1)}$. $\Psi_i^{(0)} = 0$. We iteratively approximate final Ψ_i for the ultimate rank list R^*. The expectation of stable Ψ_i is obtained when $\Psi_i^{(z)} = \Psi_i^{(z-1)}$. Final Ψ_i is expected to satisfy $\Psi_i = \alpha \mathcal{G}_i + \beta \mathcal{L}_i + \gamma \Psi_i$:

$$\Psi_i = \frac{\alpha \mathcal{G}_i + \beta \mathcal{L}_i}{1 - \gamma} = \frac{\alpha \mathcal{G}_i + \beta \mathcal{L}_i}{\alpha + \beta} \tag{10}$$

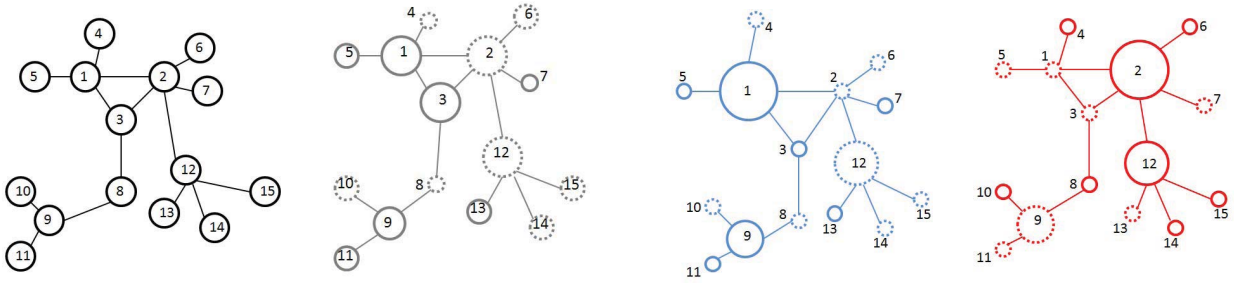
Final Ψ_i is dependent only on original global/local biased ranking scores. Equation (8) becomes more concise with no Ψ or γ: r^* is a weighted combination of global and local ranks by $\frac{\alpha}{\beta}$ ($\alpha \neq 0$, $\beta \neq 0$):

$$\begin{aligned} r_i^* &= \frac{\alpha}{\alpha + \beta} r_i^\dagger + \frac{\beta}{\alpha + \beta} r_i^\ddagger \\ &= \frac{1}{1 + \beta/\alpha} r_i^\dagger + \frac{1}{1 + \alpha/\beta} r_i^\ddagger \end{aligned} \tag{11}$$

4 Experiments and Evaluation

4.1 Datasets

There is no existing standard test set for ETTS methods. We randomly choose 6 news subjects with special coverage and handcrafted timelines by editors from 10 selected news websites: these 6 test sets consist of news datasets and golden standards to evaluate our proposed framework empirically, which amount to 10251 news articles. As shown in Table 2, three of the sources are in UK, one of them



(a) An illustrative network. (a) PageRank on t. (b) DivRank on t (c) DivRank on t'

Figure 3: An illustration of diverse ranking in a toy graph (a). Comparing (b) from general PageRank with (c),(d) from DivRank, we find a better diversity by selecting {1,9} in (c) rather than {1,3} in (b). Moreover, (c) and (d) reflect temporal biased processes on t {1,9} in (c) and t' {2,12} in (d).

is in China and the rest are in the US. We choose these sites because many of them provide timelines edited by professional editors, which serve as reference summaries. The news belongs to different categories of Rule of Interpretation (ROI) (Kumaran and Allan, 2004). More detailed statistics are in Table 3.

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.

News Subjects	#size	#docs	#stamps	#RT	AL
1.Influenza A	115026	2557	331	5	83
2.Financial Crisis	176435	2894	427	2	118
3.BP Oil Spill	63021	1468	135	6	76
4.Haiti Earthquake	12073	247	83	2	32
5.Jackson Death	37819	925	168	3	64
6.Obama Presidency	79761	2160	349	5	92

size: the whole sentence counts; #stamps: the number of timestamps; Note **average size** of subsets is calculated as: avg.size=#size/#stamps; RT: reference timelines; AL: avg. length of RT measured in sentences.

4.2 Experimental System Setups

• **Preprocessing.** As ETTS faces with much larger corpus compared with traditional MDS, we apply further data preprocessing besides stemming and stop-word removal. We extract *text snippets* representing atomic "events" from all documents with a toolkit provided by Yan et al. (2010; 2011a), by which we attempt to assign more fine-grained and accurate timestamps for every sentence within the text snippets. After the snippet extraction procedure, we filter the corpora by discarding non-event texts.

• **Compression Rate and Date Selection.** After preprocessing, we obtain numerous snippets with fine-grained timestamps, and then decompose them into temporally tagged sentences as the global collection C. We partition C according to timestamps of sentences, i.e., $C = C^1 \cup C^2 \cup \cdots \cup C^{|T|}$. Each component summary is generated from its corresponding sub-collection. The sizes of component summaries are not necessarily equal, and moreover, not all dates may be represented, so date selection is also important. We apply a simple mechanism that 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*, which indicates probable significant occurrences (Chieu and Lee, 2004). The compression rate on t_i is set as $\phi_i = \frac{|C^i|}{|C|}$.

4.3 Evaluation Metrics

The ROUGE measure is widely used for evaluation (Lin and Hovy, 2003): the DUC contests usually officially employ ROUGE for automatic summarization evaluation. In ROUGE evaluation, the summarization quality is measured by counting the number of overlapping units, such as N-gram, word sequences, and word pairs between the candidate timelines CT and the reference timelines RT. There are several kinds of ROUGE metrics, of which the most important one is ROUGE-N with 3 sub-metrics:

1 ROUGE-N-R is an N-gram recall metric:

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

2 ROUGE-N-P is an N-gram precision metric:

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

3 ROUGE-N-F is an N-gram F_1 metric:

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

I denotes a timeline. N in these metrics stands for the length of N-gram and N-gram$\in$RT denotes the N-grams in reference timelines while N-gram$\in$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 timeline and in the set of reference timelines. $\text{Count}_{\text{(N-gram)}}$ is the number of N-grams in reference timelines or candidate timelines.

According to (Lin and Hovy, 2003), among all sub-metrics, unigram-based ROUGE (ROUGE-1) has been shown to agree with human judgment most and bigram-based ROUGE (ROUGE-2) fits summarization well. We report three ROUGE F-measure scores: ROUGE-1, ROUGE-2, and ROUGE-W, where ROUGE-W is based on the weighted longest common subsequence. The weight W is set to be 1.2 in our experiments by ROUGE package (version 1.55). Intuitively, the higher the ROUGE scores, the similar the two summaries are.

4.4 Algorithms for Comparison

We implement the following widely used summarization algorithms as baseline systems. They are designed for traditional summarization without trans-temporal dimension. The first intuitive way to generate timelines by these methods is via a global summarization on collection C and then distribution of selected sentences to their source dates. The other one is via an equal summarization on all local sub-collections. For baselines, we average both intuitions as their performance scores. For fairness we conduct the same preprocessing for all baselines.

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

Centroid: The method applies MEAD algorithm (Radev et al., 2004) to extract sentences according to the following three parameters: centroid value, positional value, and first-sentence overlap.

GMDS: The graph-based MDS proposed by (Wan and Yang, 2008) first constructs a sentence connectivity graph based on cosine similarity and then selects important sentences based on the concept of eigenvector centrality.

Chieu: (Chieu and Lee, 2004) present a similar timeline system with different goals and frameworks, utilizing *interest* and *burstiness* ranking but neglecting trans-temporal news evolution.

ETTS: ETTS is an algorithm with optimized combination of global/local biased summarization.

RefTL: As we have used multiple human timelines as references, we not only provide ROUGE evaluations of the competing systems but also of the human timelines against each other, which provides a good indicator as to the upper bound ROUGE score that any system could achieve.

4.5 Overall Performance Comparison

We use a **cross validation** manner among 6 datasets, i.e., train parameters on one subject 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 4 and details are listed in Tables 4$\sim$6.

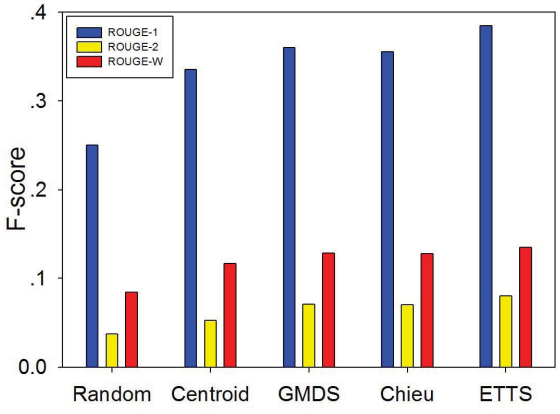


Figure 4: Overall performance on 6 datasets.

From the results, we have following observations:

• Random has the worst performance as expected.

• The results of Centroid are better than those of Random, mainly because the Centroid method takes

Table 4: Overall performance comparison on Influenza A (ROI* category: Science) and Financial Crisis (ROI category: Finance). α=0.4, kernel=Gaussian, σ=60.

	1. Influenza A			2. Financial Crisis		
Systems	R-1	R-2	R-W	R-1	R-2	R-W
RefTL	0.491	0.114	0.161	0.458	0.112	0.159
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
ETTS	**0.375**	**0.071**	**0.132**	**0.339**	**0.058**	**0.112**

Table 5: Overall performance comparison on BP Oil (ROI category: Accidents) and Haiti Quake (ROI category: Disasters). α=0.4, kernel=Gaussian, σ=30.

	3. BP Oil			4. Haiti Quake		
Systems	R-1	R-2	R-W	R-1	R-2	R-W
RefTL	0.517	0.135	0.183	0.528	0.139	0.187
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
ETTS	**0.441**	**0.107**	**0.158**	**0.436**	**0.111**	**0.145**

Table 6: Overall performance comparison on Jackson Death (ROI category: Legal Cases) and Obama Presidency (ROI category: Politics). α=0.4, kernel=Gaussian, σ=30.

	5. Jackson Death			6. Obama Presidency		
Systems	R-1	R-2	R-W	R-1	R-2	R-W
RefTL	0.482	0.113	0.161	0.495	0.115	0.163
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
ETTS	**0.358**	**0.061**	**0.130**	**0.369**	**0.074**	**0.133**

*ROI: news categorization defined by Linguistic Data Consortium.

into account positional value and first-sentence overlap, which facilitate 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 consider sentence selection from either the global or the local scope, and hence bias occurs. Mis-selected sentences result in a low recall. Generally the performance of global priority intuition (i.e. only global summarization and then distribution to temporal subsets) is better than local priority methods (only local summarization). Probable bias is enlarged by searching for worthy sentence in single dates. However, precision drops due to excessive choice of global timeline-worthy sentences.

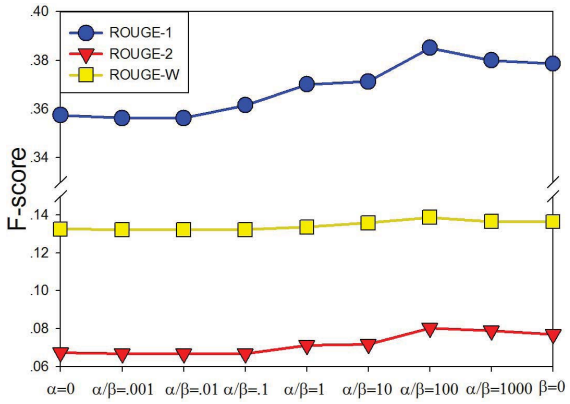


Figure 5: α/β: global/local combination.

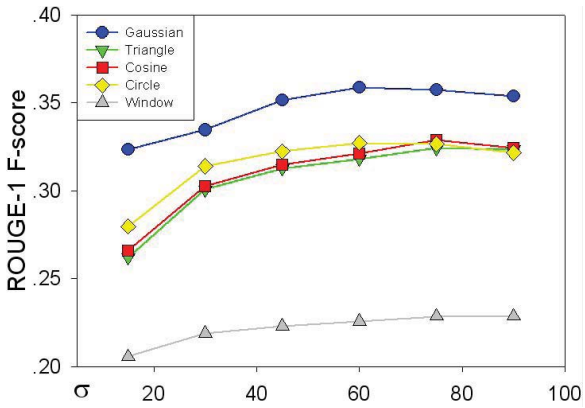


Figure 6: σ on long topics ($\geq$1 year).

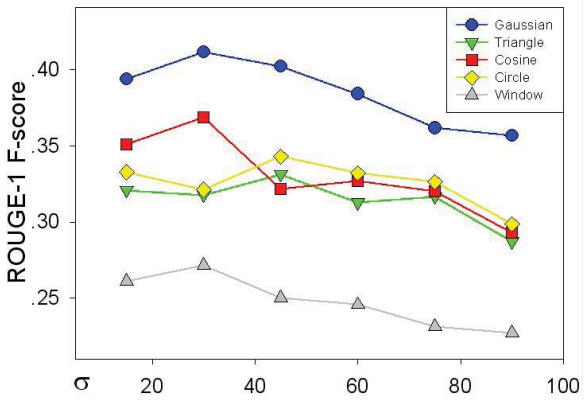


Figure 7: σ on short topics ($<$1 year).

• In general, the result of Chieu is better than Centroid but unexpectedly, worse than GMDS. The reason may be that Chieu does not capture sufficient timeline attributes. The "interest" modeled

in the algorithms actually performs flat clustering-based summarization which is proved to be less useful (Wang and Li, 2010). GMDS utilizes sentence linkage, and partly captures "correlativeness".

• ETTS under our proposed framework outperforms baselines, indicating that the properties we use for timeline generation are beneficial. We also add a direct comparison between ETTS and ETS (Yan et al., 2011b). We notice that both balanced algorithms achieve comparable performance (0.386 v.s. 0.412: a gap of 0.026 in terms of ROUGE-1), but ETTS is much faster than ETS. It is understandable that ETS refines timelines based on neighboring component summaries iteratively while for ETTS neighboring information is incorporated in temporal projection and hence there is no such procedure. Furthermore, ETS has 8 free parameters to tune while ETTS has only 2 parameters. In other words, ETTS is more simple to control.

• The performance on intensive focused news within short time range (|last timestamp−first timestamp |<1 year) is better than on long lasting news.

Having proved the effectiveness of our proposed methods, we carry the next move to identity how *global−local combination ratio* α/β and *projection kernels* take effects to enhance the quality of a summary in parameter tuning.

4.6 Parameter Tuning

Each time we tune one parameter while others are fixed. To identify how *global* and *local* biased summarization combine, we provide experiments on the performance of varying α/β in Figure 5. Results indicate that a balance between global and local biased summarization is essential for timeline generation because the performance is best when $\frac{\alpha}{\beta} \in [10, 100]$ and outperforms global and local summarization in isolation, i.e., when α=0 or $\beta = 0$ in Figure 5. Interestingly, we conclude an opposite observation compared with ETS. Different approaches might lead to different optimum of global/local combination.

Another key parameter σ measures the temporal projection influence from global collection to local collection and hence the size of neighboring sentence set. 6 datasets are classified into two groups. Subject 1, 2, 6 are grouped as long news with a time span of more than one year and the others are short news. The effect of σ varies on long news sets and short news sets. In Figure 6 σ is best around 60 and in Figure 7 it is best at about 20∼40, indicating long news has relatively wider semantic scope.

We then examine the effect of different projection kernels. Generally, Gaussian kernel outperforms others and window kernel is the worst, probably because Gaussian kernel provides the best smoothing effect with no arbitrary cutoffs. Window kernel fails to distinguish different weights of neighboring sets by temporal proximity, so its performance is as expected. Other 3 kernels are comparable.

4.7 Sample Output and Case Study

Sample output is presented in Table 7 and it shares major information similarity with the human timeline in Table 1. Besides, we notice that a dynamic ϕ_i is reasonable. Important burstiness is worthy of more attention. Fewer sentences are selected on the dates when nothing new occurs.

Interesting Findings. We notice that humans have biases to generate timelines for they have (1) preference on local occurrences and (2) different writing styles. For instance, news outlets from United States tend to summarize reactions by US government while UK websites tend to summarize British affairs. Some editors favor statistical reports while others prefer narrative style, and some timelines have detailed explanations while others are extremely concise with no more than two sentences for each entry. Our system-generated timelines have a large variance among all golden standards. Probably a new evaluation metric should be introduced to measure the quality of human generated timelines to mitigate the corresponding biases. A third interesting observation is that subjects have different volume patterns, e.g., $H1N1$ has a slow start and a bursty evolution and *BP Oil* has a bursty start and a quick decay. $Obama$ is different in nature because the report volume is temporally stable and scattered.

5 Conclusion

We present a novel solution for the important web mining problem, Evolutionary Trans-Temporal Summarization (ETTS), which generates trajectory timelines for news subjects from massive data. We formally formulate ETTS as a combination of *global* and *local* summarization, incorporating affinity and

Table 7: Selected part of timeline generated by ETTS for $BP\ Oil$.

April 20, 2010	**April 24, 2010**
s_1: An explosion on the Deepwater Horizon offshore oil drilling rig in the Gulf of Mexico, around 40 miles south east of Louisiana, causing several kills and injuries. s_2: The rig was drilling in about 5,000ft (1,525m) of water, pushing the boundaries of deepwater drilling technology. s_3: The rig is owned and operated by Transocean, a company hired by BP to carry out the drilling work. s_4: Deepwater Horizon oil rig fire leaves 11 missing.	s_1: Oil is found to be leaking from the well.
April 22, 2010	**April 26, 2010**
s_1: The US Coast Guard estimates that the rig is leaking oil at the rate of up to 8,000 barrels a day. s_2: The Deepwater Horizon sinks to the bottom of the Gulf after burning for 36 hours, raising concerns of a catastrophic oil spill. s_3: Deepwater Horizon rig sinks in 5,000ft of water.	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 23, 2010	**April 27, 2010**
s_1: The US coast guard suspends the search for missing workers, who are all presumed dead. s_2: The Coast Guard says it had no indication that oil was leaking from the well 5,000ft below the surface of the Gulf. s_3: Underwater robots try to shut valves on the blowout preventer to stop the leak, but BP abandons that failed effort two weeks later. s_4: The US Coast Guard estimates that the rig is leaking oil at the rate of up to 8,000 barrels a day. s_5: Deepwater Horizon clean-up workers fight to prevent disaster.	s_1: BP reports a rise in profits, due in large part to oil price increases, as shares rise again. s_2: The US departments of interior and homeland security announce plans for a joint investigation of the explosion and fire. s_3: Minerals Management Service (MMS) approves a plan for two relief wells. s_4: BP chairman Tony Hayward says the company will take full responsibility for the spill, paying for legitimate claims and cleanup cost.
	April 28, 2010
	s_1: The coast guard says the flow of oil is 5,000bpd, five times greater than first estimated, after a third leak is discovered. s_2: BP's attempts to repair a hydraulic leak on the blowout preventer valve are unsuccessful. s_3: BP reports that its first-quarter profits more than double to £3.65 billion following a rise in oil prices. s_4: Controlled burns begin on the giant oil slick.

diversity into a unified ranking framework. We implement a system under such framework for experiments on real web datasets to compare all approaches. Through our experiment we notice that the combination plays an important role in timeline generation, and global optimization weights slightly higher ($\alpha/\beta \in [10, 100]$), but auxiliary local information does help to enhance performance in ETTS.

Acknowledgments

This work was partially supported by NSFC with Grant No.61073082, 60933004, 70903008 and 61073081, and Xiaojun Wan was supported by NSFC with Grant No.60873155 and Beijing Nova Program (2008B03).

References

James Allan, Rahul Gupta, and Vikas Khandelwal. 2001. 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.

Hai Leong Chieu and Yoong Keok Lee. 2004. 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.

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

Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell. 1999. Summarizing text documents: sentence selection and evaluation metrics. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 121–128.

Xin Jin, Scott Spangler, Rui Ma, and Jiawei Han. 2010. Topic initiator detection on the world wide web. In *Proceedings of the 19th international conference on WWW'10*, pages 481–490.

Giridhar Kumaran and James Allan. 2004. Text classification and named entities for new event detection. In *Proceedings of the 27th annual international ACM SIGIR'04*, pages 297–304.

Chin-Yew Lin and Eduard Hovy. 2002. 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.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the Human Language Technology Conference of the NAACL'03*, pages 71–78.

Yuanhua Lv and ChengXiang Zhai. 2009. Positional language models for information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 299–306.

Qiaozhu Mei, Jian Guo, and Dragomir Radev. 2010. Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD'10*, pages 1009–1018.

R. Mihalcea and P. Tarau. 2005. A language independent algorithm for single and multiple document summarization. In *Proceedings of IJCNLP*, volume 5.

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

Russell Swan and James Allan. 2000. Automatic generation of overview timelines. In *Proceedings of the 23rd annual international ACM SIGIR'00*, pages 49–56.

Xiaojun Wan and Jianwu Yang. 2008. 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.

X. Wan, J. Yang, and J. Xiao. 2007a. Manifold-ranking based topic-focused multi-document summarization. In *Proceedings of IJCAI*, volume 7, pages 2903–2908.

X. Wan, J. Yang, and J. Xiao. 2007b. Single document summarization with document expansion. In *Proceedings of the 22nd AAAI'07*, pages 931–936.

Dingding Wang and Tao Li. 2010. 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.

Rui Yan, Yu Li, Yan Zhang, and Xiaoming Li. 2010. Event recognition from news webpages through latent ingredients extraction. In *Information Retrieval Technology - 6th Asia Information Retrieval Societies Conference, AIRS 2010*, pages 490–501.

Rui Yan, Liang Kong, Yu Li, Yan Zhang, and Xiaoming Li. 2011a. A fine-grained digestion of news webpages through event snippet extraction. In *Proceedings of the 20th international conference companion on world wide web*, WWW '11, pages 157–158.

Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang. 2011b. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *Proceedings of the 34th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '11.