Comparative Timeline Summarization via Dynamic Affinity-Preserving Random Walk

Yijun Duan and **Adam Jatowt** and **Masatoshi Yoshikawa**[1]

Abstract. Documents which contain accounts of historical events are quite common. Biographies or descriptions of entity histories like histories of places or organizations are examples of such timeline documents. Their content is explicitly or implicitly associated with timestamps indicating occurrence time of described past events. The collections of such timeline documents can be quite large and can pose challenge for readers trying to make sense of them. We then introduce a novel research task, *Comparative Timeline Summarization* (CTS), as an effective strategy to discover important similarities and differences in large collections of timeline documents for providing contrastive type of knowledge. We propose a novel summarization framework which relies on a *dynamic affinity-preserving mutually reinforced random walk* for the CTS task and evaluate it on diverse Wikipedia categories and New York Time news collections. The ROUGE evaluations demonstrate the superior performance of our method on summarizing contrastive and diverse themes over competitive baselines.

1 Introduction

Multi-Document Summarization (MDS) plays an important role in combating the problem of information overload caused by the exponential growth of documents these days, especially, ones posted on the Web. Quite many documents contain chronologically ordered content describing histories of entities or detailed accounts of past events. Biographies and history sections of Wikipedia articles are examples of such documents in which paragraphs or sentences are associated with time indicators[2] denoting the chronology of discussed events. We call such documents, *timeline documents*, or, in short, *timelines*. Note that traditional MDS techniques are not suitable for summarizing timeline documents, mainly due to their too general assumptions on sentence importance, as shown in [8, 24]. Thus they need to be adapted to properly cope with temporal character of timeline documents.

Sometimes, users would like to *compare* two collections of *timeline documents* to discover commonalities and differences between them. For example, they may be interested in questions such as *how different were the lives of French scientists from those of American scientists in the* 19^{th} *century?* or *What makes the histories of Chinese cities distinct from the ones of Japanese cities?* Note that such contrasting knowledge is difficult to be manually obtained due to relatively large number of documents that need to be analyzed, which is a time consuming work. What is more, judging the significance of encountered information requires expertise or much cognitive effort for non-professional users.

Comparative Timeline Summarization (CTS) is then a solution that we propose to automatically provide a condensed and informative document reorganization, consisting of major contrasting events chronologically ordered, for faster and better understanding of the compared sets of timeline documents. The proposed framework can be embedded into Wikipedia, search engines or other web mining services to improve users' experience when dealing with large amounts of history-related texts. For example, for a student who is interested in the comparison between histories of Chinese and Japanese cities, our proposal could greatly facilitate her understanding of these two entity sets.

What characteristics should be prominent in the CTS summary? We propose three major ones. The first is *Coverage*. Summary sentences should be important and cover majority of information in the input document collections. The second one is *Distinctness*. The summary should deliver the major differences between compared timeline collections by extracting the most discriminative sentences in each document group. The last one is *Diversity*. The information overlap among summary sentences should be as minimal as possible due to the length restriction. The summary should thus cover diverse aspects of comparison.

The problem of CTS is not trivial. It faces the challenges of (1) modeling the three objectives and (2) generating summaries which take into account the significance of the *temporal dimension*, while satisfying the above-described three objectives. To address the above challenges we make the following observations w.r.t timeline documents: (1) timeline documents can often be divided into latent time units representing cohesive atomic time units (e.g., "eras" in city histories), where sentences within the same latent time unit tend to exhibit high coherence and similarity. (2) Sentences at different time units are not assumed to be completely isolated due to the evolving characteristics of timeline documents (e.g., "consequences" or "follow-ups"). Based on these observations we propose computing two kinds of sentence importance:

- *Local importance*. A sentence is *locally important* in a given time unit if it is semantically similar to sentences in its document set, while being semantically dissimilar to sentences of the contrasting document set in that time unit.
- *Global importance*. A sentence is *globally important* if it is *locally important* in many different time units. It can be inferred that a summary consisting of many globally important sentences will naturally satisfy our three constraints.

To effectively discover locally important sentences, we propose the *affinity-preserving mutually reinforced Markov random walk* model (APMRRW) (see Sec. 2.1). Different from the classical Markov ran-

[1] Graduate School of Informatics, Kyoto University, Japan, email: {yijun, adam}@dl.kuis.kyoto-u.ac.jp, yoshikawa@i.kyoto-u.ac.jp

[2] Typically, temporal expressions serve as such indicators. They can be however implicit when time is obvious from context.

dom walk model based on democratic normalization, APMRRW normalizes the transition matrix by its first norm to preserve the original affinity relations between sentences, which leads to the amplification of the effect of locally important candidates, as well as the suppression of bad sentences. We derive the ranking for APMRRW based on its quasi-stationary distribution (see Theorem 1), and prove its equivalence to the spectral relaxation for the Integer Quadratic Programming (IQP) formulation of the classical graph matching problem [13] (see Theorem 2). Furthermore, the *dynamic affinity-preserving mutually reinforced random walk* (D-APMRRW) is proposed to identify globally important sentences (see Sec. 2.2). It reweights transition matrix during each local summarization, to equip the surfer with knowledge about what a diverse summary should be.

We test our approaches on 12 manually annotated datasets including diverse Wikipedia categories and New York Time news collections in comparison to competitive baselines of various types. Our experimental results show that contrastive themes can be successfully summarized from two comparable timeline document sets, as measured using ROUGE metric.

To sum up, we make the following contributions in this paper:

1. We introduce a new research task of *comparative timeline summarization* to summarize contrastive knowledge from two sets of compared timeline documents.
2. We propose a novel *affinity-preserving mutually reinforced random walk* model which preserves the original affinity relations between sentences. The preservation of affinity results in a local summary consists of more salient and discriminative information.
3. We design a *dynamic affinity-preserving mutually reinforced random walk* to produce the global summary, which adjusts transition matrix for improving the diversity of summary sentences in the random walk process.
4. Experiments on diverse Wikipedia categories and the New York Time corpus show competitive performance of our method.

2 Proposed Method

In this section, we present the key components of our summarization framework. We first propose in Sec. 2.1 an *affinity-preserving mutually reinforced random walk* model (APMRRW) to locally score sentences. We then present in Sec. 2.2 the *dynamic affinity-preserving mutually reinforced random walk* model (D-APMRRW) which flexibly allows for generating globally important and diverse summary.

2.1 Affinity-Preserving Mutually Reinforced Random Walk

We now describe the formulation of APMRRW for locally scoring sentences. Intuitively, locally salient sentences in a given time unit are assumed to be similar to sentences in the same set while dissimilar to sentences in the contrasting set, within the same time unit.

Given two input sets of timeline documents D_A and D_B, we construct a two-layer graph $G = (V_A, V_B, E_{AA}, E_{BB}, E_{AB})$ in the following steps:

- For each sentence s_i^A in D_A and each sentence s_i^B in D_B, we create a normal vertex v_i^A in layer L_A and a normal vertex v_i^B in layer L_B, respectively.
- For each pair of vertices v_i and v_j of the *same* layer, we create a bi-directed edge (v_i, v_j) between v_i and v_j. Moreover, we associate it with a weight w_{ij}^s indicating the *similarity* between v_i and v_j, which is computed as

$$w_{ij}^s = sim_{cosine}(v_i, v_j) \tag{1}$$

- For each pair of vertices v_i and v_j of the *different* layers, we create a bi-directed edge (v_i, v_j) between v_i and v_j. Each edge is associated with a weight w_{ij}^d indicating the *difference* between v_i and v_j, which is computed as

$$w_{ij}^d = 1 - sim_{cosine}(v_i, v_j) \tag{2}$$

- For each layer $L \in \{A, B\}$ an additional absorbing vertex v_0^L is created. Each absorbing vertex is self-transitioned, which means there are no edges coming from it.
- For each pair of a normal vertex v_i and an absorbing vertex v_0 of the *same* layer, we create a uni-directed edge from v_i to v_0.
- For each pair of a normal vertex v_i in layer $L1$ and an absorbing vertex v_0 of the *different* layer $L2$, we create a uni-directed edge from v_i to v_0.

Here, $V_A = \{v_0^A, v_1^A, v_2^A, ..., v_m^A\}$ and $V_B = \{v_0^B, v_1^B, v_2^B, ..., v_n^B\}$ denote the sets of vertices contained in layer L_A and L_B, respectively. v_0^A and v_0^B denote the *absorbing vertex* in each layer while the others are normal vertices. The illustration of a graph for APMRRW is shown in Fig. 1.

We then construct affinity metrics W_{AA}, W_{BB}, W_{AB} and W_{BA}. Specifically, we have $W_{AA} = [w_{ij}^s | v_i^A, v_j^A]$, $W_{BB} = [w_{ij}^s | v_i^B, v_j^B]$, $W_{AB} = [w_{ij}^d | v_i^A, v_j^B]$ and $W_{BA} = [w_{ij}^d | v_i^B, v_j^A]$ computed by Equations (1) and (2). Now, motivated by affinity-preserving random walk [23], we construct transition matrices P_{AA}, P_{BB}, P_{AB} and P_{BA}, which are normalized by the *first norm* of their correspondent affinity metric. The goal of such normalization is to enable random walk process to distinguish between good and bad vertices. For instance, P_{AA} is formulated as

$$P_{AA} = \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{e} - W_{AA} \cdot \mathbf{e}/||W_{AA}||_1 & W_{AA}/||W_{AA}||_1 \end{pmatrix} \tag{3}$$

Here, $\mathbf{0}$ and $\mathbf{e}$ are $m \times 1$ vectors with all elements 0 and 1, respectively. Metrics P_{BB}, P_{AB} and P_{BA} are defined in a similar way. It can be observed that these transition metrices are *"soft"* stochastic metrics, where *"soft"* means that the sum of row elements in the matrix can be less than 1, and the leakage represents the amount of tendency for a random walker to be absorbed. It is small for a vertex having a large affinity with other vertices, and large for a bad candidate vertex. Then the *affinity-preserving mutually reinforced random walk* on graph G is defined as:

Definition 2.1. ***Affinity-Preserving Mutually Reinforced Random Walk.*** *The discrete-time Markov chain* $\mathbf{X} = \{(\mathbf{X}_A^0, \mathbf{X}_B^0)^T, (\mathbf{X}_A^1, \mathbf{X}_B^1)^T, (\mathbf{X}_A^2, \mathbf{X}_B^2)^T ...\}$ *with state space and transition probability matrix* $\boldsymbol{Q}$*, where* $\boldsymbol{Q}$ *is given by*

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_B \end{pmatrix} \tag{4}$$

Here, $(\mathbf{X}_A^t, \mathbf{X}_B^t)$ denotes the scores of vertex set V_A and that of vertex set V_B at the t-th iteration, where $\mathbf{Q}_A = P_{AA}P_{AB}P_{BB}P_{BA}$ and $\mathbf{Q}_B = P_{BB}P_{BA}P_{AA}P_{AB}$.

Based on $\mathbf{Q}$, $(\mathbf{X}_A^t, \mathbf{X}_B^t)$ integrates the *within-layer* and *mutually-reinforced between-layer* propagation. However, the steady state distribution $(\mathbf{X}_A^*, \mathbf{X}_B^*)$ is always $\{\{1, \mathbf{0}\}^T, \{1, \mathbf{0}\}^T\}$. To guarantee a good characterization of the sentence ranking distribution on graph G, we adopt the quasi-stationary distribution [5, 6] as the distribution of unabsorbed random walkers. Then for numerical computation of the sentence scores, we can iteratively run Eq. (5) until convergence,

in the similar way as PageRank [16]:

$$(\bar{\mathbf{X}}_A^{t+1}, \bar{\mathbf{X}}_B^{t+1})^T = \frac{d \cdot \mathbf{Q}^{'}(\bar{\mathbf{X}}_A^t, \bar{\mathbf{X}}_B^t)^T + (1-d) \cdot (\bar{\mathbf{X}}_A^0, \bar{\mathbf{X}}_B^0)^T}{\left\| d \cdot \mathbf{Q}^{'}(\bar{\mathbf{X}}_A^t, \bar{\mathbf{X}}_B^t)^T + (1-d) \cdot (\bar{\mathbf{X}}_A^0, \bar{\mathbf{X}}_B^0)^T \right\|_1} \tag{5}$$

where d is the damping factor that trades off between the transition specified by $\mathbf{Q}^{'}$ and the teleport vector $(\bar{\mathbf{X}}_A^0, \bar{\mathbf{X}}_B^0)^T$. Let $\Theta = ||\mathbf{W_{AA}}||_1||\mathbf{W_{AB}}||_1||\mathbf{W_{BB}}||_1||\mathbf{W_{BA}}||_1$, $\mathbf{Q}^{'}$ is computed as

$$\mathbf{Q}^{'} = \frac{1}{\Theta} \cdot \begin{pmatrix} \mathbf{W_{AA}W_{AB}W_{BB}W_{BA}} & \mathbf{0} \\ \mathbf{0} & \mathbf{W_{BB}W_{BA}W_{AA}W_{AB}} \end{pmatrix} \tag{6}$$

Note that in order to guarantee the convergence of eq. (5), $(\bar{\mathbf{X}}_A, \bar{\mathbf{X}}_B)$ is normalized after each iteration such that the scores sum to 1.

Alternatively, to compute the quasi-stationary vertex ranking distribution in APMRRW, we state the following theorem

Theorem 1. *The closed-form solution $\bar{\mathbf{X}}_L^*, L \in \{A, B\}$ is the normalized principal eigenvector of χ_L when damping factor $d = 1$, where $\chi_A = W_{AA}W_{AB}W_{BB}W_{BA}$ and $\chi_B = W_{BB}W_{BA}W_{AA}W_{AB}$.*

Proof. Let $\Delta = ||W_{AA}||_1||W_{AB}||_1||W_{BB}||_1||W_{BA}||_1$. In stationary distribution, $(\bar{\mathbf{X}}_A^{*(t+1)}, \bar{\mathbf{X}}_B^{*(t+1)})^T = (\bar{\mathbf{X}}_A^{*(t)}, \bar{\mathbf{X}}_B^{*(t)})^T$ holds. Following the definition of quasi-stationary distribution, we then have

$$\begin{aligned}
(\bar{\mathbf{X}}_A^{*(t+1)}, \bar{\mathbf{X}}_B^{*(t+1)})^T &= (\frac{\mathbf{X}_A^{*(t+1)}[1:m]}{1 - \mathbf{X}_A^{*(t+1)}[0]}, \frac{\mathbf{X}_B^{*(t+1)}[1:n]}{1 - \mathbf{X}_B^{*(t+1)}[0]})^T \\
&= (\frac{\chi_A \mathbf{X}_A^{*(t)}[1:m]/\Delta}{1 - (\mathbf{X}_A^{*(t)}[0] + (\mathbf{e} - \chi_A \mathbf{e}/\Delta) \cdot \mathbf{X}_A^{*(t)}[1:m])}, \\
&\quad \frac{\chi_B \mathbf{X}_B^{*(t)}[1:n]/\Delta}{1 - ((\mathbf{X}_B^{*(t)}[0] + (\mathbf{e} - \chi_B \mathbf{e}/\Delta) \cdot \mathbf{X}_B^{*(t)}[1:n]))})^T \\
&= (\frac{\chi_A \mathbf{X}_A^{*(t)}[1:m]/\Delta}{1 - (1 - \chi_A \mathbf{e} \mathbf{X}_A^{*(t)}[1:m]/\Delta)}, \\
&\quad \frac{\chi_B \mathbf{X}_B^{*(t)}[1:n]/\Delta}{1 - (1 - \chi_B \mathbf{e} \mathbf{X}_B^{*(t)}[1:n]/\Delta)})^T \\
&= (\frac{\chi_A \mathbf{X}_A^{*(t)}[1:m]}{\chi_A \mathbf{e} \mathbf{X}_A^{*(t)}[1:m]}, \frac{\chi_B \mathbf{X}_B^{*(t)}[1:n]}{\chi_B \mathbf{e} \mathbf{X}_B^{*(t)}[1:n]})^T \\
&= (\frac{\chi_A \bar{\mathbf{X}}_A^{*(t)}}{\chi_A \mathbf{e} \bar{\mathbf{X}}_A^{*(t)}}, \frac{\chi_B \bar{\mathbf{X}}_B^{*(t)}}{\chi_B \mathbf{e} \bar{\mathbf{X}}_B^{*(t)}})^T \\
&= (\bar{\mathbf{X}}_A^{*(t)}, \bar{\mathbf{X}}_B^{*(t)})^T
\end{aligned} \tag{7}$$

Thus $(\bar{\mathbf{X}}_A^*, \bar{\mathbf{X}}_B^*)^T$ must satisfy $(\lambda_A \bar{\mathbf{X}}_A^*, \lambda_B \bar{\mathbf{X}}_B^*)^T = (\chi_A \bar{\mathbf{X}}_A^*, \chi_B \bar{\mathbf{X}}_B^*)^T$. According to the *Perron Frobenius theorem* [17, 10], given irreducible non-negative matrices (χ_A, χ_B) and stochastic vectors $(\bar{\mathbf{X}}_A^*, \bar{\mathbf{X}}_B^*)$, $\bar{\mathbf{X}}_A^*$ and $\bar{\mathbf{X}}_B^*$ are the normalized Perron-Frobenius eigenvector of χ_A and χ_B, respectively, corresponding to their maximal eigenvalue. □

2.2 Dynamic Affinity-Preserving Mutually Reinforced Random Walk

We now present the idea of the D-APMRRW model (see Fig. 2) for generating *globally* important sentences, which, as mentioned before, are assumed to be locally important in many time units. At each time unit t, sentences of two compared document sets within t are locally scored. Based on the ranking scores, a particular number of top sentences are selected as the summary at the time unit t, based on the aforementioned APMRRW model. Such locally generated summary

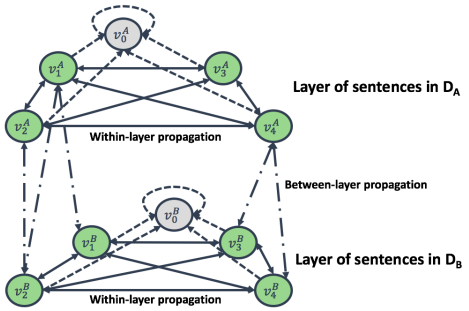


Figure 1: An illustration of two-layer APMRRW. V_A and V_B denote the sets of sentences contained in D_A and D_B, respectively, and v_0^A and v_0^B are absorbing vertices playing a role of soaking unreliable ranking scores from bad sentences and of distinguishing good sentences. Three different types of edges corresponding to different relations: V_A-to-V_A, V_B-to-V_B, V_A-to-V_B are present, where the first two within-layer relations are based on node similarity and the last between-layer relation is based on node dissimilarity.

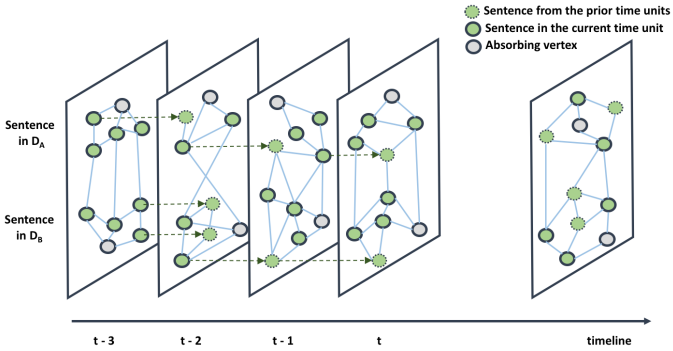


Figure 2: An illustration of D-APMRRW model. At each time unit t (except for the start point), candidate sentences S consist of local sentences in t (denoted by solid green circles) and prior summaries from $t-1$ (denoted by dash green circles). Then the aforementioned APMRRW (see Sec. 2.1) is applied on S at time unit t and the generated summary will be passed over to $t+1$. During the summarization in t, old sentences from $t-1$ are normalized "locally" (i.e., there are no links between them to the absorbing vertex), while the new sentences in t are normalized "globally" (i.e., unreliable ranking scores of new sentences will be absorbed). Such operation of reweighting transition matrix is proposed to improve the diversity of final summary.

is next passed over to the following time unit as "past" information. Then a new set of sentences to be summarized consisting of both the previous summary and sentences in the new unit is constructed, and the same local summarization procedure is applied. Thus, another local summary is generated and is passed over to the next time point. Like this, the APMRRW is repeatedly applied to the candidate sentences at each time unit. The summarization process works dynamically, and a comparative summary can be generated for any length of period, and at any level of granularity of time units.

However, we notice that APMRRW does not guarantee the diversity constraint of summary in the random walk process. This characteristic is embodied in the theorem below.

Theorem 2. *The quasi-stationary distribution $(\bar{\mathbf{X}}_A^*, \bar{\mathbf{X}}_B^*)^T$ of affinity-preserving mutually reinforced random walk is proportional to the solution $\tilde{\mathbf{x}}$ of below quadratic score function when damping factor $d = 1$*

$$\begin{aligned} &\arg\max(\tilde{\mathbf{x}}^T \mathbf{\Omega} \tilde{\mathbf{x}}) \\ &s.t.\ \|\tilde{\mathbf{x}}\|_2 = 1 \end{aligned} \tag{8}$$

where $\mathbf{\Omega}$ *is given by*

$$\mathbf{\Omega} = \begin{pmatrix} \chi_A & \mathbf{0} \\ \mathbf{0} & \chi_B \end{pmatrix} \tag{9}$$

Proof. For real matrix $\mathbf{\Omega}$ and non-zero vector $\tilde{\mathbf{x}}$ which satisfy $\|\tilde{\mathbf{x}}\|_2 = 1$, we maximize $\tilde{\mathbf{x}}^T\mathbf{\Omega}\tilde{\mathbf{x}}$ subject to this constraint by using a Lagrange multiplier: $\tilde{\mathbf{x}}^T\mathbf{\Omega}\tilde{\mathbf{x}} + \omega \cdot (\|\tilde{\mathbf{x}}\|_2 - 1)$, and differentiating with respect to the components of $\tilde{\mathbf{x}}$. Then we obtain the equation $\mathbf{\Omega}\tilde{\mathbf{x}} - \omega\tilde{\mathbf{x}} = 0$, so the extrema are precisely the eigenvectors of $\mathbf{\Omega}$. If $\tilde{\mathbf{x}}$ is an eigenvector, then it follows immediately that the value of $\tilde{\mathbf{x}}^T\mathbf{\Omega}\tilde{\mathbf{x}}$ is the corresponding eigenvalue λ_i. Thus $\tilde{\mathbf{x}}^T\mathbf{\Omega}\tilde{\mathbf{x}}$ is maximal when $\tilde{\mathbf{x}}$ is corresponding to the principal eigenvector of $\mathbf{\Omega}$.

On the other hand, solution $(\bar{\mathbf{X}}_A^*, \bar{\mathbf{X}}_B^*)^T$ is the normalized principal eigenvector of $\mathbf{\Omega}$. Thus we have $(\bar{\mathbf{X}}_A^*, \bar{\mathbf{X}}_B^*)^T = \frac{\tilde{\mathbf{x}}}{\|\tilde{\mathbf{x}}\|_1}$, hence $(\bar{\mathbf{X}}_A^*, \bar{\mathbf{X}}_B^*)^T$ is proportional to $\tilde{\mathbf{x}}$. □

It can be observed from the above theorem that APMRRW tends to attain a stationary distribution which *maximizes the total sum of affinity* (i.e., the similarity for sentences in the same document set and dissimilarity for sentences in different sets) between sentences in the summary. To equip the surfer with knowledge about what a diverse summary should be, we propose to *reweight transition matrix* at each time unit during the passing process. The key idea is that, sentences from the summary of the previous time unit (which are already good and diverse) should be normalized *locally*, while local sentences (which are a mixture of many bad sentences and few good sentences) should be normalized *globally*.

More concretely, at each time unit t, let S^{t-1} and D^t denote the previous summary and the local sentences, respectively. Suppose sentences in S^{t-1} are already salient and diversified, thus we encourage the surfer to explore more in the neighborhood of S^{t-1} rather than end in the absorbing vertex. Based on this consideration, if sentence s_i is included in S^{t-1}, each element in the corresponding i-th row of affinity matrix W will be normalized "locally" by the sum of row elements (i.e., democratic normalization), so that a surfer at s_i will never be absorbed. Otherwise, those elements will be normalized "globally" by the maximum sum of row elements in W (i.e., normalization by first norm). By differentiating the normalization methods for different sentences, the saliency and diversity constraints of a produced summary will be naturally highlighted.

3 Experimental setup

3.1 Research questions

We first list the research questions that guide our experiments:
RQ1: How does our dynamic affinity-preserving mutually reinforced random walk model perform on comparative timeline summarization? Does it outperform baselines? (see Sec. 4.1)
RQ2: Is the affinity-preserving mechanism helpful for extracting locally important sentences? Is the dynamic ranking framework helpful for identifying globally important sentences? Does the operation of reweighting transition matrix facilitate the diversity of summary? (see Sec. 4.2)
RQ3: How does our model perform w.r.t. different values of damping factor d and the length of time unit l? What is the optimal value of d and l? (see Sec. 4.3)

3.2 Datasets

We employ 12 datasets in our experiments which belong to two types. Both types have been used in the previous works [8, 19, 7, 3, 25]. The basic statistics about our datasets are shown in Tab. 1. In total, 27,251 documents are used in experiments.

The first type [3, 8] consists of diverse Wikipedia categories and lists, including *locations*, *persons* and *organizations*. To facilitate the evaluation, we used existing Wikipedia categories and lists of moderate size, with which all the annotators were quite familiar. These are the histories of 3 pairs of Wikipedia categories including location categories, (Japanese cities vs. Chinese cities), organization categories (western teams of NBA league in North America vs. eastern teams of NBA) and person categories (Japanese Prime Ministers till the end of WW2 vs. Japanese Prime Ministers after WW2), respectively.

The second type of datasets we use consist of news articles selected from the New York Times corpus [25], which is a collection of 99,872 articles published by the New York Times between January 1990 and July 2016. Each news article is assigned to "section" such as *Business*, *Sports*. In our experiments, we focus on the comparison of 3 pairs of comparable news article collections, where each collection consists of news associated with the same section over time. Specifically, these compared sections are *U.S.* vs. *World* , *Science* vs. *Technology* and *Arts* vs. *Fashion & Style*.

Table 1: Summary of the Wikipedia and news datasets.

General Description	# Docs	# Sentences
Japanese Cities (D_A^1)	532	22,045
Chinese Cities (D_B^1)	357	6,444
Western NBA Teams (D_A^2)	15	3,755
Eastern NBA Teams (D_B^2)	15	3,701
Japanese PMs pre WW2 (D_B^3)	32	2,338
Japanese PMs post WW2 (D_B^3)	30	1,715
U.S. News (D_A^4)	7,541	616,628
World News (D_B^4)	6,013	404,944
Science News (D_A^5)	587	35,686
Technology News (D_B^5)	1,161	73,484
Arts News (D_A^6)	7,388	563,108
Fashion & Style News (D_B^6)	3,580	264,990

3.3 Reference Summary

To the best of our knowledge, there are no human-made comparative summaries for our task. In the field of text summarization (particularly in special settings, domains or for special applications), researches need to rely on self-built benchmark datasets (e.g., [22, 12, 7]). We have then hired five human judges who are not authors of this paper to manually annotate the experimental datasets. The annotators were asked to write up to 300-words long[3] *reference summary* for each document set that will help in grasping the contrastive content of the input document collections. In particular, after we pooled the summaries created by all the analyzed methods, the annotators were asked to conduct the following two data annotation tasks: (1) *Task 1*. The first task was to highlight all the representative and discriminative sentences in the pool to form the salient sentences set. There was no limit imposed on the number of highlighted sentences. The annotators did not know which systems generated which summaries. (2) *Task 2*. The second task was to write up to 300-words long *reference summary* of the text selected in the first task. During the two tasks, the annotators were allowed to utilize any external resources or search engines to verify the correctness of the results.

3.4 Analyzed Methods

Type 1. We first test the performance of our proposed models. We prepare D-APMRRW for the overall process as described in Sec. 2.

[3] The average length of an English sentence is around 15 words [22], so we choose 300 as the number of words for the summary size of 20 sentences.

We also test D-MRRW and APMRRW as the models that skip the affinity-preserving in Sec. 2.1, and the dynamic ranking in Sec. 2.2, respectively. Similarly, we test D-APMRRW* as the model that does not adjust transition matrix at each time unit in Sec. 2.2.

Type 2. We then make comparisons with two popular comparative summarization models (denoted as CS models): the discriminative sentence selection model (DSS, [22]) and the integer linear programming model (ILP, [12]).

Type 3. The third type of strategy performs two commonly used multi-document summarization methods (denoted as MDS models) as baselines: (1) LexRank [9] that ranks sentences via a Markov random walk strategy and (2) ClusterCMRW [21] which scores sentences by a clustering-based approach.

Type 4. Finally, three state-of-the-art timeline summarization approaches (denoted as TS models) which rely on an exemplar-based Markov random walk model (E-MRW, [8]), an evolutionary timeline summarization model (ETS, [24]), and an online graph-based model (OGM, [20]) are used for evaluation.

3.5 Experimental Settings

In this study, we set the summary size to 20 sentences, following [8, 7]. To represent terms and sentences, we adopt the commonly-used Skip-gram model [15]. We obtain the distributed vector representations of each word by training the Skip-gram model on the entire English Wikipedia from 2016 using the gensim Python library [18]. The vector representation of a sentence is a TF-IDF weighted combination of the vectors of terms. The number of dimensions of word vectors is experimentally set to 200.

3.6 Evaluation Metrics

To assess the saliency of summaries, we evaluate all the models with the following measures:

ROUGE-1.5.5 toolkit [14]. The ROUGE is a widely used metric which has been officially adopted by DUC for automatic summarization evaluation. In the experiments, we report the f-measure values of ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-W and ROUGE-SU.

Precision. The ROUGE measures mainly reflect the recall. We further examine the rate of summary sentences included in the human-labeled important sentence set of the dataset as follows:

$$precision = \frac{\{summary\_sentences\} \cap \{labeled\_sentences\}}{\{summary\_sentences\}} \tag{10}$$

To evaluate the diversity among the set of summary sentences S we compute the *diversity* as follows:

$$diversity = \frac{1}{|S|^2} \sum_{s_i \in S} \sum_{s_j \in S} 1 - sim_{cosine}(s_i, s_j) \tag{11}$$

4 Results and Discussion

4.1 Overall Performance

We start by addressing **RQ1** and analyze whether D-APMRRW is effective for the comparative timeline summarization task. First, Tab. 2 summarizes the ROUGE performance of our models compared to the one of 7 baselines. Performance on Wikipedia categories and on New York Time datasets is reported in columns 2-6 and in columns 7-11, respectively. Tab. 3 lists the precision and diversity scores for all the analyzed methods on both the dataset types.

From Tab. 2, we find that D-APMRRW achieves the best performance in terms of all ROUGE metrics, except for ROUGE-SU on New York Times datasets. Specifically, when compared with *Type 1* (DSS and ILP), *Type 2* (LexRank and ClusterCMRW) and *Type 3* (E-MRW, ETS and OGM) methods, it outperforms them by 91.1%, 53.1% and 58.0% on the Wikipedia datasets, and by 74.1%, 36.7% and 47.8% on the New York Times datasets in terms of ROUGE-1 score, respectively. From Tab. 3, D-APMRRW achieves the highest *precision* score on the Wikipedia categories, and the highest *diversity* score on the New York Times datasets, respectively.

Generally, it can be observed that the baseline methods perform relatively poorly. For multi-document summarization methods, the plausible reason can be that they neglect the significant temporal dimension and evolutionary characteristics of timeline documents, nor do they incorporate distinct information into summarization. For example, Lexrank only tends to select very general sentences which are similar to many other sentences in the entire document set. Contrastingly, the comparative summarization methods conceptually focus more on discovering sentences that deliver set-specific information. For example, DSS uses a multivariate normal generative model to extract content which best describe the unique characteristics of each document group. Such procedure can prevent them from embodying historic significance, which to some extent may explain its second-worst performance among all baselines. Finally, the timeline summarization methods exhibit relatively competitive performance. However, they also suffer from the ignorance of discriminative information, which may interest annotators when producing reference summary.

4.2 Assumption Validation

To answer **RQ2** regarding the components of our proposal, Tab. 4 shows the results of the proposed method's variants when evaluating by ROUGE, and Tab. 5 displays their results when evaluating by *precision* and *diversity*. From Tab. 4, we find that D-APMRRW, which considers the overall process as described in Sec. 2, has the best ROUGE scores among all its variants on both the Wikipedia datasets and the New York Time datasets. From Tab. 5, we can see that it also achieves the best *precision* and *diversity* scores.

More concretely, when compared to APMRRW, D-MRRW and D-APMRRW*, D-APMRRW shows a 47.3%, 23.3% and 22.2% increase in terms of ROUGE-1 score, and a 65.2%, 31.9% and 21.3% increase in terms of ROUGE-2 score, respectively. In terms of *accuracy*, DPMRRW outperforms other variants with 64.5% of an average increase. When looking at *diversity*, D-APMRRW offers an increase over D-APMRRW* (which does not conduct the operation of reweighting transition matrix) of up to 26.2%.

On the other hand, APMRRW gets the lowest score at all evaluation metrics, whereas D-MRRW exhibits a similar performance with D-MRRW*. Again, D-APMRRW and all its variants tend to obtain higher scores on the New York Times datasets than on the Wikipedia ones.

Hence, we conclude that the proposed assumptions all help to improve the quality of generated summaries, and that dynamic ranking (Sec. 2.2) offers a more significant performance increase than the local affinity preservation (Sec. 2.1) and the reweighting of transition matrix (Sec. 2.2). This observation implies that globally important sentences are better candidates for summary than the locally important ones. Specially, we demonstrate that reweighting transition matrix at each time unit during dynamic ranking indeed moves to the sentence distribution that induces a more diversified summary.

Table 2: (**RQ1**) ROUGE performance of all analyzed summarization models.

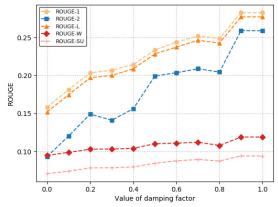
	Wikipedia Categories					New York Times				
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU
DSS [22]	0.130	0.057	0.050	0.052	0.022	0.164	0.083	0.077	0.100	0.042
ILP [12]	0.140	0.063	0.089	0.060	0.028	0.191	0.095	0.129	0.108	0.065
LexRank [9]	0.159	0.076	0.156	0.072	0.040	0.214	0.107	0.202	0.115	0.071
ClusterCMRW [21]	0.178	0.093	0.175	0.077	0.044	0.238	0.136	0.214	0.128	0.076
E-MRW [8]	0.184	0.104	0.181	0.076	0.043	0.229	0.134	0.220	0.124	0.078
ETS [24]	0.191	0.112	0.185	0.078	0.050	0.271	0.174	0.240	0.139	**0.082**
OGM [20]	0.115	0.051	0.104	0.062	0.036	0.127	0.075	0.113	0.078	0.040
D-APMRRW	**0.258**	**0.154**	**0.208**	**0.089**	**0.060**	**0.309**	**0.193**	**0.248**	**0.151**	0.078

Table 3: (**RQ1**) Precision and Diversity scores of all models.

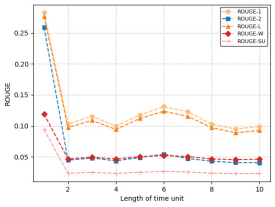
	Wikipedia Categories		New York Times	
	Precision	Diversity	Precision	Diversity
DSS [22]	0.000	0.207	0.000	0.231
ILP [12]	0.067	0.185	0.033	0.212
LexRank [9]	0.067	0.169	0.100	0.198
ClusterCMRW [21]	0.067	0.224	**0.167**	0.209
E-MRW [8]	0.033	0.183	**0.167**	0.194
ETS [24]	0.100	**0.234**	0.133	0.228
OGM [20]	0.067	0.176	0.100	0.216
D-APMRRW	**0.150**	0.194	0.133	**0.239**

4.3 Parameter Tuning

Turning to **RQ3**, Fig. 3a and Fig. 3b show the ROUGE performance of D-APMRRW w.r.t. the value of damping factor d and the length of time unit l used in dynamic ranking, respectively. We first test d within the range [0, 1] and with a step of 0.1. In Fig. 3a, we can see that the value of d has an effect on the performance of summarization. The ROUGE scores peak when d equals 0.9; the performance basically keeps increasing until $d = 0.9$, yet it decreases after d exceeds 0.9. This observation is consistent with the commonly used value of damping factor equal to 0.85 in the literature [4]. When it comes to the length of time unit, we change it in the range [1,10] with a step of 1 year. When the time unit is larger than 1, the system achieves worse performance due to the plausible reason that many good candidate sentences are discarded during the dynamic ranking, as we enlarge the candidate sentences set while keeping the summary size fixed per local scoring.



Rouge Performance w.r.t. damping factor d.



Rouge Performance w.r.t. time unit l.

Figure 4: (RQ3) Parameter tuning.

4.4 Example Summary

We present in this section the comparative summary of contrastive themes between a typical history of major 357 Chinese cities and that of main 532 Japanese cities (see Sec. 3.2), generated by our method D-APMRRW. The summary consists of two timelines, each containing 10 events ordered chronologically, as shown in Fig. 5. Our model produces summaries in which each event is in the form of a sentence from the history of a particular entity. However, to facilitate readers' understanding of summarized contrastive themes, we choose to generalize the top-scored summary sentences to output the set of descriptive words representing in a general way a given event group (theme) based on the method proposed in [8], as shown in Tab. 6 and Tab. 7. We manually assign labels based on the words representing each event group.

The summary describes some import comparisons between the history of Chinese cities and that of Japanese cities. For example, different from Chinese cities, it can be observed that Japanese cities frequently suffered from *Natural Disasters* such as earthquakes, tsunamis and typhoons (e.g. the *Hanshin Earthquake* in 1994), and Japan is paying particular attention to *Nuclear* issues (e.g. the *Fukushima Daichi Nuclear Disaster* in 2011). In addition, since a long time ago in the history of Japanese cities, *Castles* (e.g. the famous *Himeji Castle*) were continued to be built as an important symbol of centres of governance, and various Japanese *Festivals* with local customs have been popular (e.g. the *Gion Matsuri*). They are both typical representatives of Japanese unique culture. On the other hand, there are more records related to the more ancient time in the history of Chinese cities, such as *Spring and Autumn Period*, *Tang Dynasty* and *Ming Dynasty*. The *Revolution* in the early 20th century and the *Reform* around 1978 are two key turning points in the social development of Chinese cities. The modern *Education* in China started from the middle 20th century; at the same time China's population started to grow rapidly, embodied in the event *Population*. In addition, it can be observed that Chinese cities hosted many *Sport* events (e.g. the *AFC Asian Cup* in 2004 and the *Peking Olympics* in 2008) in the 2000's. As we can see, most of the comparisons are clear and convey comparative historical knowledge.

5 Related Work

Comparative Summarization. Comparative summarization requires providing short summaries from multiple comparative aspects. Wang *et al*. [22] propose a discriminative sentence selection method based on a multivariate normal generative model aiming to extract sentences that are best describing the unique characteristics of each document group. Huang *et al*. [12] formulate the task of comparative news summarization as an optimization problem of selecting sentences to maximize the score of comparative and representative evidences based on an integer linear programming (ILP) model. Ren *et al*. [19] explicitly consider contrast, relevance and diversity for summarizing contrastive themes by adopting a hierarchical nonparametric Bayesian model to infer hierarchical relations among topics for enhancing the diversity of themes. Recently, differential topic models have also been used to measure sentence discriminative capability for comparative summarization [11].

Time Summarization. Timeline Summarization defined as the summarization of sequences of documents (typically, news articles about the same event) has been actively studied in the recent years. In [2], Alonso *et al*. present a timeline generation design that captures the most salient events along with the most popular keywords as annotations alongside a timeline. Yan *et al*. [24] propose the evo-

Table 4: (**RQ2**) ROUGE performance of all variants of proposed model.

	Wikipedia Categories					New York Times				
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU
APMRRW	0.180	0.092	0.169	0.077	0.043	0.205	0.118	0.195	0.105	0.063
D-MRRW	0.218	0.123	0.191	0.085	0.049	0.242	0.140	0.228	0.130	0.067
D-APMRRW*	0.209	0.128	0.188	0.081	0.047	0.255	0.158	0.234	0.136	0.072
D-APMRRW	**0.258**	**0.154**	**0.208**	**0.089**	**0.060**	**0.309**	**0.193**	**0.248**	**0.151**	**0.078**

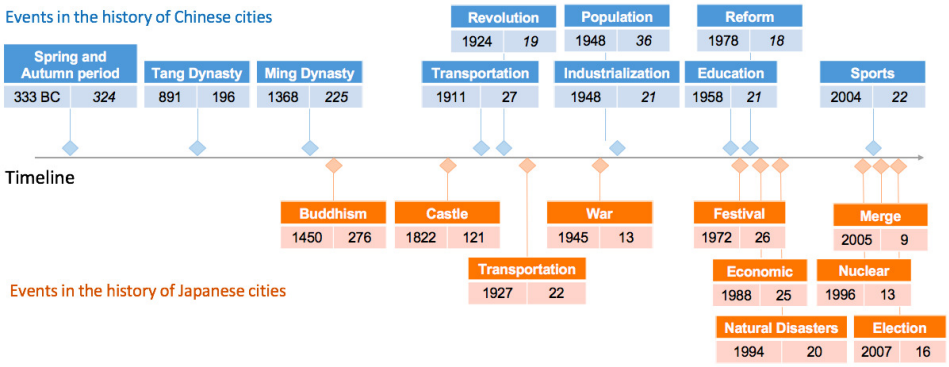


Figure 5: The summary of 20 comparative themes between a typical history of Chinese cities learned from 357 instances and that of Japanese cities learned from 532 instances. Each event is illustrated by a manually created label based on data from Tab. 6 and Tab. 7, along with its median (left value) and the standard deviation of occurrence time (right value).

Table 5: (**RQ2**) ROUGE performance of all model variants.

	Wikipedia Categories		New York Times	
	Precision	Diversity	Precision	Diversity
APMRRW	0.067	0.155	0.033	0.172
D-MRRW	0.100	0.162	0.083	0.195
D-APMRRW*	0.100	0.152	**0.133**	0.191
D-APMRRW	**0.150**	**0.194**	**0.133**	**0.239**

Table 6: Events in Chinese cities summary. Due to space limit we show 5 events, and for each event we show up to top 10 descriptive words.

Event	Terms
Spring and Autumn period	qin, warring, chu, dynasty, capital conquered, zhou, subjugated, county, vassal
Revolution	communist, rebellion, nationalist, revolt, army kmt, war, rebel, party, revolution
Industrialization	company, steel, iron, plant, installed oil, factory, production, cotton, mine
Population	population, million, per, estimated, urban reached, tripled, exceeded, xpcc, increased
Reform	development, economic, growth, industry, investment port, zone, bank, billion, reform

Table 7: Events in Japanese cities summary. Due to space limit we show 5 events, and for each event we show up to top 10 descriptive words.

Event	Terms
Buddhism	temple, period, shrine, year, buddhist history, area, site, built, nara
War	war, air, world, raid, army japanese, bombing, naval, base, imperial
Economic	billion, gdp, population, million, employment city, industry, greater, increase, economy
Natural Disasters	earthquake, tsunami, damage, suffered, typhoon caused, struck, magnitude, killed, city
Nuclear	city, nuclear, evacuee, accident, fukushima student, public, problem, caused, rapid

lutionary timeline summarization (ETS) to compute evolution timelines consisting of a series of time-stamped summaries. David *et al*. [3] present a method for discovering biographical structures based on a probabilistic latent variable model. Their approach summarizes timestamped biographies to a set of event classes along with the typical times when those events occur. Duan *et al*. [8] propose a summarization task aimed at generating gists of histories of multiple entities. Satoko *et al*. [20] present a graph-based algorithm for online summarization of time-series documents. Abdalghani *et al*. [1] address the problem of identifying important events in the past, present, and future from semantically-annotated large-scale document collections.

To the best of our knowledge, we are the first to work on comparative summarization of timeline documents. Unlike in the case of the general comparative summarization tasks, the input documents to our task have strong temporal characteristics that need to be considered. On the other hand, in contrast to timeline summarization tasks, we aim to discover discriminative and contrasting information for comparing the sets of timeline documents.

6 Conclusion

This work introduces a special kind of summarization task - *Comparative Timeline Summarization* (CTS) and proposes effective approaches towards solving it. The unique character of our proposed summarization allows capturing important comparative aspects of evolutionary trajactories hidden in two sets of timeline documents. We approach the CTS task by applying the *dynamic affinity-preserving mutually reinforced random walk* model which is capable of generating globally important and diverse summary. We have shown that the proposed model outperform various competitive baselines in the experiments on 6 pairs of manually annotated datasets using ROUGE toolkit. In future, we will use abstractive summarization strategies for increasing the readability of generated summaries.

7 Acknowledgements

This research has been partially supported by MEXT Kakenhi grants (#17H01828, #18K19841 and #19H04215).

REFERENCES

[1] Abdalghani Abujabal and Klaus Berberich, 'Important events in the past, present, and future', in *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pp. 1315–1320, New York, NY, USA, (2015). ACM.

[2] Omar Alonso and Kyle Shiells, 'Timelines as summaries of popular scheduled events', in *Proceedings of the 22nd international conference on world wide web*, pp. 1037–1044. ACM, (2013).

[3] David Bamman and Noah A Smith, 'Unsupervised discovery of biographical structure from text', *Transactions of the Association for Computational Linguistics*, **2**, 363–376, (2014).

[4] Sergey Brin and Lawrence Page, 'Reprint of: The anatomy of a large-scale hypertextual web search engine', *Computer networks*, **56**(18), 3825–3833, (2012).

[5] Minsu Cho, Jungmin Lee, and Kyoung Mu Lee, 'Reweighted random walks for graph matching', in *European conference on Computer vision*, pp. 492–505. Springer, (2010).

[6] John N Darroch and Eugene Seneta, 'On quasi-stationary distributions in absorbing discrete-time finite markov chains', *Journal of Applied Probability*, **2**(1), 88–100, (1965).

[7] Yijun Duan and Adam Jatowt, 'Across-time comparative summarization of news articles', in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pp. 735–743, New York, NY, USA, (2019). ACM.

[8] Yijun Duan, Adam Jatowt, and Katsumi Tanaka, 'Discovering typical histories of entities by multi-timeline summarization', in *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pp. 105–114. ACM, (2017).

[9] Günes Erkan and Dragomir R Radev, 'Lexrank: Graph-based lexical centrality as salience in text summarization', volume 22, pp. 457–479, (2004).

[10] Georg Frobenius, Ferdinand Georg Frobenius, Ferdinand Georg Frobenius, Ferdinand Georg Frobenius, and Germany Mathematician, 'Über matrizen aus nicht negativen elementen', (1912).

[11] Lei He, Wei Li, and Hai Zhuge, 'Exploring differential topic models for comparative summarization of scientific papers.', in *COLING*, pp. 1028–1038, (2016).

[12] Xiaojiang Huang, Xiaojun Wan, and Jianguo Xiao, 'Comparative news summarization using linear programming', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*, pp. 648–653. Association for Computational Linguistics, (2011).

[13] Marius Leordeanu and Martial Hebert, 'A spectral technique for correspondence problems using pairwise constraints', in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pp. 1482–1489. IEEE, (2005).

[14] Chin-Yew Lin and Eduard Hovy, 'Automatic evaluation of summaries using n-gram co-occurrence statistics', in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 71–78. Association for Computational Linguistics, (2003).

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781*, (2013).

[16] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, 'The pagerank citation ranking: Bringing order to the web.', Technical report, Stanford InfoLab, (1999).

[17] Oskar Perron, 'Zur theorie der matrices', *Mathematische Annalen*, **64**(2), 248–263, (1907).

[18] Radim Řehůřek and Petr Sojka, 'Software Framework for Topic Modelling with Large Corpora', in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta, (May 2010). ELRA. http://is.muni.cz/publication/884893/en.

[19] Zhaochun Ren and Maarten de Rijke, 'Summarizing contrastive themes via hierarchical non-parametric processes', in *SIGIR*, pp. 93–102. ACM, (2015).

[20] Satoko Suzuki and Ichiro Kobayashi, 'On-line summarization of time-series documents using a graph-based algorithm', in *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, (2014).

[21] Xiaojun Wan and Jianwu 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*, pp. 299–306. ACM, (2008).

[22] Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong, 'Comparative document summarization via discriminative sentence selection', *TKDD*, **6**(3), 12, (2012).

[23] Kexiang Wang, Tianyu Liu, Zhifang Sui, and Baobao Chang, 'Affinity-preserving random walk for multi-document summarization', in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 210–220, (2017).

[24] Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang, 'Evolutionary timeline summarization: a balanced optimization framework via iterative substitution', in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 745–754. ACM, (2011).

[25] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong, 'Dynamic word embeddings for evolving semantic discovery', in *WSDM*, pp. 673–681. ACM, (2018).