

# Leveraging Aging theory in topic-focused multi-document timeline summarization

No Author Given

No Institute Given

## 1 Abstract

Topic-focused multi-document summarization plays an important role in helping readers to get the main information from any topic. Many approaches are proposed to generate the timeline summarization, but seldom consider the life circle of each topic. In this paper, we leveraging aging theory to present the sentence feature, and train the classification model with the SMOTEBoost technology. We evaluate our approach on two corpus, one of which is a public data set, the other one is our manual annotation data set. Experiment results show that our method can improve the timeline summarization significantly.

## 2 Introduction

Everyday thousands of news stories reporting different events are published on the Internet. These reports are disordered and people have to read most of them to know what is happening which is a time-consuming job undoubtedly. How can we get useful information about an event efficiently? Automatic summarization has been such a method solving this kind of information overloading since Luhn [?] proposed it in 1958. And numerous pages have been published in this field, ranging from single document to multiple documents, from extraction to abstraction, from traditional document to web document, email, blog and other types of genre. However, these research work focus on the central idea of document or document set ignoring the temporal characteristics of events. As a result, people cannot catch the changes of events over time efficiently.

Recent years, topic detection and tracking (TDT) which detects new events from the large scale news stream and tracks them as events going on draws researchers' attention. But it did not display events properly, and people still have to read all the relevant reports to get what they want to know about the event. However, we are still enlightened by its usage of tracking which make us decide to generate a timeline summary consisting of a series of individual small summaries with sentences both important and diverse to help people understand the progress of an event more quickly.

Every event goes through a life cycle of birth, growth, maturity and death, which means that special terms utilized for describing different events experience a similar life cycle. Aging theory [?] is a model exploited in event detection task which tracks life cycles of events using energy function. The energy of an

event increases when the event becomes popular, and it diminishes with time. In our opinion, it can also be used for summarization to help us find out the daily hot terms of events. Then people can obtain what new changes happen as events going on.

One challenge lie in compute the importance of sentences, which is decided by terms occurring at the documents in keywords-based summarization. But different authors use different words to express the same meaning and lots of synonyms. In order to find the core word in the news without the influence caused by the synonym and polysemy, we use latent semantic analysis (LSA) [?] to handle the dataset. LSA is a robust unsupervised technique for deriving an implicit representation of text semantics based on observed co-occurrence of words to find semantic units of news.

The other challenge is how to handle the sentences. In each topic, only a few sentences will be labled as the summarization sentence, that is, the data set is imbalanced. This situation caused a problem that the training model will prefer the normal sentence. SMOTEBoost [?] combine SMOTE [?] and boost technology in order to improve the precscie for minority class through resampling. This has been proved effective for imbalanced data set.

In this paper, we generate news event timeline summary by considering both temporal and semantic characteristics. We first extract the features from five aspects to represent each sentence. Then, classification model is build with SMOTEBoost. Last, we choose sentences from candidates to from the summary and display them with timeline, so that people can track the progress of event easily and quickly.

The remainder of this paper is organized as follows: Section 2 reviews some related works on summarization. We discuss our approach about how to leverage aging theory to gain sentence feature and train the classification model with SMOTEBoost technology in section 3. Our experiments and some discusses are described in section 4. Section 5 presents our conclusions and some future plans.

### 3 Related Work

**Topic-focused multi-document summarization(TMS)** aims at gain main information from multiple text about the same topic. There are two ways to achive this goal, one is extract important sentences, the other one is build new sentence to express the key idea. In this paper, we focuse on the former method.

One of the most popular extractive multi-document summarization method is MEAD [?], which take term frequency, sentence position, first-sentence overlap to present the feature of each sentece. Xiaojun Wan [?] proposed a extractive approach based on manifold-ranking about the information richness and novelty.

**Timeline summarization(TS)** gain enough attraction with the development of Topic Track and Detection(TDT). Lots of timeline summarization methods have been developed recently. ETS [?] formulate the task as an optimization problem via iterative substitution from a set of sentences with four requirements.

Giang Binh Tran[?] investigate five different sentence feature and leverage SVM-Rank to optimize the summarization task. Wayne Xin Zhao [?] take social attention involved to compute the importance. ETTS [?] utilize trans-temporal characteristics to gain the summary.

**Aging theory** proved to be effective to track what stage of life cycle for news. Chen et al. [?] [?] applied this to model the news event's life cycle and utilized the concept of energy to track it. Because our aim is to gain the summary of multi-documents of news domain, we consider aging theory is worth using to extract the feature of sentence.

## 4 Our Approach

### 4.1 Key Concepts

**Topic-focused:** What we value most is an event grouped from several web news articles, such as the "the missing of the malaysia airlines plane" from BBC. These articles show us the cause, the progress and the results about the topic. Most of our summarization technology's application scenario is working for TDT system. **Timeline Summaries:** Generally speaking, timeline is one of display form for the summaries. But timeline summaries should show us the progress of this topic not just display according to the time sequence. Under this condition of the requirement, timeline summary of each day should describe the most important thing happened in that day.

We give the formal definition of topic-focused multi-document timeline summarization(TMTS) as follows:

**Input:** Given a set of documents  $D = \{d_1, d_2, \dots, d_n\}$  which should cover progress of the topic in the time span  $T = \{t_1, t_2, \dots, t_m\}$ . We segment each document to sentences and group them by the date to form sentences  $S = \{s_1, s_2, \dots, s_m\}$ .

**Output:** The TMTS should output the summaries along the date and each summary is the main idea of what occurred in that day, i.e.  $O = \{o_1, o_2, \dots, o_m\}$ , where  $o_i$  means the summary of sentences from all the documents of that day  $s_i$ .

### 4.2 Sentence Feature Selection

In order to represent sentence, we extract five kinds of features as follows:

**Surface feature:** this contains features computed by basic statistics, such as the length of sentence, the counts of noun words and stop words, the position in this document and paragraph, and whether it contains person name or not.

**Importance feature:** this feature aim to represent the importance of this sentence. The weight of sentence is computed through linear combination of term weights with latent semantic analysis.

**Aging feature:** We use this feature to show the life cycle of this sentence.

**Noviety feature:**

**Topic feature:**

### 4.3 Model Trainning

With the help of labled data, we convert this summarization task to pairwise classification problem. The positive data is sentences labeld to summary, otherwise is negative.

Because the count of summary sentence is much less than the normal sentence, the train data set is unblanced. In order to reduce this reflect, SMOTE-Boost method is used to train the classification model.

## 5 Evaluation

### 5.1 Evaluation metric

Rogue  
Prccesiction

### 5.2 Experiment on public data set

The public data set form Giang Binh Tran[?] is used in this research.

### 5.3 Experiment on muannl labeld data

## 6 Future