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 also 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 the 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 development 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 descripting 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.

The importance of sentences is decided by terms occurring at the documents in keywords-based summarization. But different authors use different words to express a same meaning and many words has several meanings. So identifying the implicit semantics of news can improve summary quality greatly. Here, we propose an incremental model based on latent semantic analysis (LSA) [**?**] which 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.

As descripted above, in this paper, we generate news event summary by considering both temporal and semantic characteristics. We first utilize the aging theory [**?**] to extract hot terms from news which reports the same event according to their energy during the time interval we choose (i.e. one day). Then we identify the semantic units of news with the incremental latent semantic analysis model. Last, we construct a semantic text relationship map, choose sentences which are both important and novel to form the summary and display them using a timeline so that people can track event trajectory easily and quickly.

The remainder of this paper is organized as follows: Section 2 reviews some related works on summarization. We discuss our approach of event timeline summarization using aging theory and incremental latent semantic analysis model 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 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. Giang Binh Tran[**?**] investigate five different sentence feature and leverage SVMRank to optimize the summarization task.

Timeline summarization is

Aging thory

4 Our Approach

4.1 Sentence Feature Selection

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

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

Importance feature: this feature aim to respesent the impantance of this sentence. The weight of sentence is computed through linear combination of term weights with latem senmatic analysis.

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

Noviety feature:

Topic feature:

4.2 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, SMOTEBoost method is used to train the classification model.

5 Evaluation

5.1 Evaluation metric

Rogue

Prcesesiction

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