Cascaded Regression Analysis Based Temporal Multi-document Summarization

Ruifang He, Bing Qin, Ting Liu, Sheng Li
Information Retrieval Lab, Harbin Institute of Technology P.O.Box 321, HIT,P.R.China 150001
E-mail: rfhe@ir.hit.edu.cn, http://ir.hit.edu.cn/

Keywords: temporal multi-document summarization, temporal semantic labeling, macro importance discriminative model, micro importance discriminative model

Received: February 5, 2009

Temporal multi-document summarization (TMDS) aims to capture evolving information of a single topic over time and produce a summary delivering the main information content. This paper presents a cascaded regression analysis based macro-micro importance discriminative model for the content selection of TMDS, which mines the temporal characteristics at different levels of topical detail in order to provide the cue for extracting the important content. Temporally evolving data can be treated as dynamic objects that have changing content over time. Firstly, we extract important time points with macro importance discriminative model, then extract important sentences in these time points with micro importance discriminative model. Macro and micro importance discriminative models are combined to form a cascaded regression analysis approach. The summary is made up of the important sentences evolving over time. Experiments on five Chinese datasets demonstrate the encouraging performance of the proposed approach, but the problem is far from solved.

Povzetek: Metoda kaskadne regresije je uporabljena za izdelavo zbirnega besedila.

1 Introduction

Multi-document summarization is a technology of information compression, which is largely an outgrowth of the late twentieth-century ability to gather large collections of unstructured information on-line. The explosion of the World Wide Web has brought a vast amount of information, and thus created a demand for new ways of managing changing information. Multi-document summarization is the process of automatically producing a summary delivering the main information content from a set of documents about an explicit or implicit topic, which helps to acquire information efficiently. It has drawn much attention in recent years and is valuable in many applications, such as intelligence gathering, hand-held devices and aids for the handicapped.

Temporal multi-document summarization (TMDS) is the natural extension of multi-document summarization, which captures evolving information of a single topic over time. The greatest difference from traditional multi-document summarization is that it deals with the dynamic collection about a topic changing over time. It is assumed that a user has access to a stream of news stories that are on the same topic, but that the stream flows rapidly enough that no one has the time to look at every story. In this situation, a person would prefer to dive into the details that include the most important, evolving concepts within the topic and have a trend analysis.

The key problem of summarization is how to identify important content and remove redundant content. The common problem for summarization is that the information in different documents inevitably overlaps with each other, and therefore effective summarization methods are needed to contrast their similarities and differences. However, the above application scenarios, where the objects to be summarized face to some special topics and evolve with time, raise new challenges to traditional summarization algorithms. One challenge for TMDS is that the information in summary must contain the evolving content. So we need to effectively take into account this temporally evolving characteristics during the summarization process. Thus a good TMDS must include information as much as possible, keeping information as novel as possible. In this paper, we focus on how to summarize the series news reports by the generic and extractive way.

Considering the temporal characteristic of the series news reports at different levels of topic detail, redundancy is a good feature. We adopted cascaded regression analysis to model the temporal redundancy from the macro and micro view. We hierarchically extract important information with the macro and micro importance discriminative models. We detected the important time points based on macro importance discriminative model, and extracted the important sentences based on micro importance discriminative model. Macro and micro importance discriminative models are combined to form a cascaded regression analysis model. This method not only reduces the complexity of the problem, but also fully mines the temporal characteristics of evolving data over time. The summary is made up of

the important sentences evolving over time. Experiments on five Chinese datasets demonstrate the encouraging performance of the proposed approach, but the problem is far from solved.

The rest of this paper is organized as follows: Section 2 introduces related work. The details of the proposed approach are described in Section 3. Section 4 presents and discusses the evaluation results. We conclude this paper and discuss future work in Section 5.

2 Related work

Temporal summary is a relatively new research direction, which originates from text summarization and topic detection and tracking (TDT). It is also related to time line construction techniques. Alan et al.[1] firstly put forward the concept of temporal summary inspired by TDT in SIGIR2001. Given a sequence of news reports on certain topic, they extract useful and novel sentences to monitor the changes over time. Usefulness is captured by considering whether a sentence can be generated by a language model created from the sentences seen to date. Novelty is captured by comparing a sentence with prior sentences. They report that it is difficult to combine the two factors successfully. Other researchers exploit distribution of events and extract the hot topics on time line by statistical measures. Swan and Allan[8] employ χ^2 statistics to measure the strength that a term is associated with a specified date, and then extract and group important terms to generate “topics" defined by TDT. In [3], Chen et al. import the aging theory to measure the “hotness" of a topic by analyzing the temporal characteristic of news report. The aging theory implies that a news event can be considered as a life form that goes through a life cycle of birth, growth, decay, and death, reflecting its popularity over time. Then hot topics are selected according to energy function defined by aging theory. Lim et al.[5] anchor documents on time line by the publication dates, and then extract sentences from each document based on surface features. Sentence weight is adjusted by local high frequency words in each time slot and global high frequency words from all topic sentences. They evaluate the system on Korean documents and report that time can help to raise the percentage of model sentences contained in machine generated summaries. Jatowt and Ishizuka[2] investigate the approaches to monitor the trends of dynamic web documents. They employ a simple regression analysis on word frequency and time to identify whether terms are popular and active. The importance of a term is measured by its slope, intercept and variance. The weight of a sentence is measured by the sum of the weights of the terms inside the sentence. The sentences with highest scores are extracted into a summary. However, they do not report any quantitative evaluation results. In [7], Mani is devoted to temporal information extraction, knowledge representation and reasoning, and try to apply them to multi-document summarization. In [4], Li et al. explore whether the temporal distribution information helps to enhance event-based summarization based on corpus of DUC2001.

Due to different tasks, the above researches do not uniformly incorporate the temporal characteristics. While macro and micro importance discriminative models based on cascaded regression analysis approach can mine the temporal characteristics at different levels of topic detail and produce summary.

3 Cascaded regression analysis approach

We know news has strong temporal characteristic. If there is a novel happening, many websites will concern it, and naturally produce vast relevant news reports, this time point would be very important at the moment. If happening is gradually disappearing, the relevant new reports will also decrease accordingly, whose importance would reduce. From macro view, redundancy implies the whole trends of the happening, which also hints the more finer progress from micro view. We model the temporal redundancy from macro and micro view, and integrate macro and micro importance discriminative models into a cascaded regression analysis framework for the content selection for TMDS. Algorithm 1 shows the basic steps.

Algorithm 1 Framework for Temporal Multi-document Summarization based on Cascaded Regression Analysis

Input: Stream of Chinese series news stories within the same topic; Output: Sentences containing important events evolving over time;

1: Parse the documents into the set of the sentences, and recognize and resolve the time expressions contained in sentences;
2: Construct the article/sentence count-time distribution curve;
3: Convert the curve given by step 2 to be the relative importance curve of the time points;
4: Extract the important time points with macro importance discriminative model;
5: Extract important events with micro importance discriminative model in each important time point;
6: Rank sentences according to the publication time and the real time;

3.1 Selection of content unit

Since series news reports are made up of time points, each time point consists of articles/sentences. Here, the ith time point t_i, the jth sentence s_j is formalized as the following, respectively: $t_i = \{s_j\}$,
$s_j = \{T_p, T_r, Trigger, Scope\}, i, j = 1...n$, Tp is publication time, Tr is real time through time resolution,

$Trigger$ is the set of trigger words, and $Scope$ is the sentence description containing events. Reference to the definition of event in ACE evaluation[1], a trigger word indicates the existence of an event. However, Chinese event extraction technology is not mature, we ignore the relevant attributes of event, including type, subtype, modality, polarity, genericity and tense. Generally, trigger word is verb or action noun. We just consider the situation of verb so as to simplify the question. Thus, the j th sentence can also be simply formalized as follows: $s_j = \{v_k\}, j, k = 1...n$.

The importance of a sentence depends on the importance of the verbs contained in a sentence. Based on the above analysis, we choose the time point and verb as the content unit of importance discrimination from macro and micro view, respectively.

3.2 Macro importance discriminative model

In the new World Wide Web environment, the number of news articles is increasing dramatically, and we can conveniently and instantaneously get the rich data. Usually, the reports about the same story from the different websites are mostly similar, especially the start and end time points and the important time points. With the evolution of an abrupt news story, the number of the news articles or events will form a distribution curve along the time axis. From the intuitive observation of macro view, this temporal characteristic of news articles gives us a good illumination on TMDS.

For example, figure 1 shows the temporal trends about the number of the news reports on the topic *Solomon turbulence* from the Sina. The horizontal axis is time, and the vertical axis is the number of articles or sentences. For the benefit of clearly observing the temporal characteristics of the curve and comparing the difference in extracting the important time points between articles and sentence count-time distribution curve, we convert it into the relative importance curve of time point and enlarge 100 times. This transformation can further help us modify the choice of the important time points. The concrete method is as follows: the relative importance value of time point is computed by the ratio of the article/sentence count in each time point to the article/sentence count in the highest peak. Figure 2 shows the curve through transformation.

According to this kind of distribution curve and our intuition, we give the macro importance discriminative model, including one assumption and one definition.

Assumption 1: The start and end time points, and the time points having more documents contain important information with a high probability. Valleys and slowly changing time points contain unimportant information with a high probability.

Definition of Slowly changing time point: If the left slope and right slope of the current time point are both lower than (we empirically assign 2 to λ), we say this time

[1] ACE2007 evaluation plan: http://projects.ldc.upenn.edu/ace/intro.html

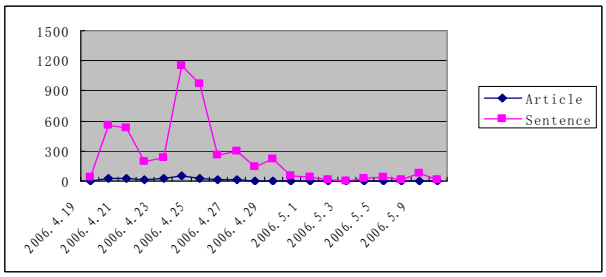

Figure 1: Article/Sentence count-time distribution about the Solomon turbulence news report.

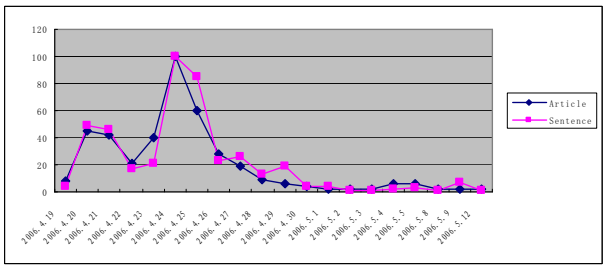

Figure 2: Relative importance distribution curve of time points.

point changes slowly. They are defined as follows:

$$\text{left\_Slope}(t_i) = \frac{\text{RI(left)}}{\text{Td(left)}} \quad (1)$$

$$\text{right\_Slope}(t_i) = \frac{\text{RI(right)}}{\text{Td(right)}} \quad (2)$$

$\text{RIleft} = \text{I}(\text{t}_\text{i}) - \text{I}(\text{t}_{\text{i}-1})$, $\text{RI(right)} = \text{I}(\text{t}_{\text{i}+1}) - \text{I}(\text{t}_\text{i})$,$\text{Td(left)} = \text{date}(\text{t}_\text{i}) - \text{date}(\text{t}_{\text{i}-1})$, $\text{Td(right)} = \text{date}(\text{t}_{\text{i}+1}) - \text{date}(\text{t}_\text{i})$, $\text{i} = 2...\text{n}$, $\text{right\_Slope}(t_i)$, $\text{right\_Slope}(t_i)$ are formed by the current time point, the left adjacent one, and the right adjacent one. RI(left), RI(right) is the relative importance difference value computed by the current time point, the left adjacent one, and right adjacent one, respectively. $\text{I}(\text{t}_\text{i})$ is the relative importance value of the ith time point. Td(left), Td(right) is the duration that the left one, the right one deviates from the current one, respectively. $\text{date}(\text{t}_\text{i})$ is the date of the ith time point.

Based on the above description, we give the algorithm of detecting the important time points with importance discriminative model.

3.3 Micro importance discriminative model

In order to extract the important sentences, we need to define the importance scoring scheme. Trigger words are core representatives of event, whose importance can reflect the importance of sentence. Therefore, we statistically analyze the importance of trigger words and define three kinds of scoring schemes as follows:

Algorithm 2 Detecting the important time points

Input: all time points; Output: the importance points;

1: Use climbing algorithm to find all the peaks and valleys, and keep the start and end time points;
2: Remove the valleys and the slowly changing time points, and get four time point sets figured out by the two kinds of relative importance distribution curve of time points;
3: Compute the intersection of four time point sets, then get the important time point set;

TFIOF based scoring scheme Depending on the basic idea of the feature weight about TFIDF, TFIOF is proposed to compute the importance of a trigger. $\forall t_i, i, k = 1...n$,

$$tfiof(v_k) = tf_i(v_k) \times log\frac{n}{of(v_k)} \quad (3)$$

$tf_i(v_k)$ is the occurrence number of v_k in the current time point. n is the number of all time points, $of(v_k)$ is the number of time points containing v_k.

Slope based scoring scheme $\forall t_i, t_{i-1}, i = 2...n, k = 1...n$,

$$left\_Slope(v_k) = \frac{tf_i(v_k) - tf_{i-1}(v_k)}{Td(left)} \quad (4)$$

$tf_{i-1}(v_k)$ denotes the occurrence number of v_k in the left adjacent time point. $left\_Slope(v_k)$ is the instantaneous left slope of the event on behalf of v_k, which adopts the linear regression method to express the consequent relation between time and event variables. If it is a positive value, it means that the event triggered by v_k is emerging, or the past event triggered by v_k is disappearing. It also shows whether this trigger word is active in the local scope.

Variance analysis based scoring scheme The model for the arrival of trigger words can be considered as a random process, and the arrival of every trigger word is a random variable. The variance value of the random variable X on behalf of v_k is represented as $Variance(X) = E\{(X - E(X))^2\}$ It represents the average magnitude of X in term of importance for a period of time, which helps us to detect the important trigger words in the global scope. The larger the $Variance(X)$ is, the more precious the trigger word. Every trigger word has a variance value. In order to compare the importance of trigger words in each time point, we normalize every random variable as X^*:

$$X^* = \frac{X - E(X)}{\sqrt{Variance(X)}} \quad (5)$$

Based on the importance scoring schemes, we give the algorithm of the important sentences extraction and ranking, see algorithm 3.

Algorithm 3 Selection and ranking of sentences

Input: Stream of series news reports through preprocessing; Output: Assume s_i to be the set of final summary in t_i, initially $s_i = \phi$;

1: Extract the trigger words;
2: Compute the importance weight of each trigger word according to three scoring schemes;
3: Rank the trigger words according to the weight from step 2, respectively;
4: **repeat**
5: For each time point, according to each ranked v_k, select the most important, but not redundant sentence including v_k with highest weight sum to add into s_i
6: **if** $\mathrm{v_k} \subset \mathrm{s_i}$ **then**
7: Compute the value of the goal function: $\mathrm{f(v_k)} = \frac{|\mathrm{s_i} \cup \mathrm{s_j}|}{\mathrm{N}} - \frac{|\mathrm{s_i}|}{\mathrm{N}}$
8: **end if**
9: **if** $\mathrm{f(v_k)} \leq \lambda(\lambda = 2/\mathrm{N})$ **then**
10: s_j is the redundant sentence, where N is the number of trigger words in the ith time point
11: **end if**
12: **until** Summary length is satisfied
13: Rank the sentences within different time points by their publication date, and rank the sentences within the time points by their real date; If the two sentences have the same real date, we donot care their relative rank;

4 Experiments

4.1 Corpus and evaluation metrics

TMDS is a new research, and there is no public corpus and evaluation metric. Therefore we have to build the corpus and the evaluation metric.

Corpus Our Chinese corpus construction includes two parts, one is the construction of raw corpus, another is the construction of reference summary. Five groups of Chinese data set are chosen from Sina's[2]international news topics between 2005 and 2006. Table1 illustrates the settings of the corpus, where there are five topics, 78 time points, 734 articles, 13486 sentences. Simultaneously, for each date set, we let experts annotate three groups of reference summary in term of the compression rate 10% and 20%.

ID	#time points	#articles	#sentences
1	20	214	4310
2	25	250	5253
3	3	17	101
4	20	158	2278
5	10	95	1544

Table 1: Corpus settings

Evaluation Metrics ROUGE[6] is used as the evalua-

[2]http://news.sina.com.cn

tion metric, which has been widely adopted by DUC for automatic summarization evaluation. It measured summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the candidate summary and the reference summary. ROUGE toolkit reported separate scores for 1, 2, 3 and 4-gram, and also for longest common subsequence co-occurrences and so on. However, this evaluation metrics faces to English. Based on it, we develop a Chinese-style ROUGE-N evaluation tool 'CROUGE-N'. The evaluation measure is F score:

$$F = \frac{\sum_{i=1}^{i=n} F_i}{n}, F_i = \frac{2P_iR_i}{P_i + R_i} \quad (6)$$

$i = 1...n$, i denotes the number of reference summary. P_i, R_i, F_i is Precision, Recall and F score, respectively.

4.2 Experiment results and analysis

Two groups of experiments are designed to validate the performance of hierarchical regression analysis approach for TMDS.

Experiment1: Micro importance discriminative model

In the first experiment, we use the micro importance discriminative model to produce the summaries under compression rate (CRate) 10% and 20%.

CRate	TFIOF	Slope	Variance
10%	17.29%	19.11%	14.81%
20%	27.18%	27.56%	27.02%

Table 2: Performance of content selection from micro view with CROUGE-4

The goal of Experiment 1 is to evaluate the performance of different micro importance discriminative modelsčňwhich extracts the important content from the fine particle scale. Table 2 shows that slope based micro model has the best performance whatever compression rate is 10% or 20%. It is because that this model can better represent the instantaneously temporal characteristics of series news reports. The performance of variance based micro model is the lowest, however, it is still meaningful. Variance of random variable indicates the average changing magnitude of its importance. It can capture the importance information from the global scope. While TFIOF based micro model is easy to be implemented and can extract important information from local and global scope. Three models observe micro information from different views, respectively. We can adopt the different micro models according to different practical applications. The more effective model incorporating their merits will be explored in the future. When the compression rate is 20%, the performance difference between three models is little. It shows that more ordinary sentences are added into the summary, while our models is apt to capture the particular sentences.

Experiment2: Macro and micro importance discriminative model

Based on the best micro importance discriminative model adopting slope scoring scheme, we further validated the performance of hierarchically extracting important information with macro and micro importance discriminative model.

The experiment results from Table 3 validated that macro and micro importance discriminative model displays the better system performance than the single micro model. The linear regression based macro and micro importance discriminative model that we adopted receives the best performance. Macro model is used to extract the important time points, which helps to have a coarse content selection. In the whole process, we try to mine the temporal characteristic of the articles, events and terms from the macro and micro view, and use the regression analysis to summarize the relationship between the time and the frequency of articles, sentences and terms. Macro importance discriminative model and micro importance discriminative model have the recursive properties to some extent. No matter what the slope used to select important time points or the slope used to extract important trigger words, their slope value from the regression different from zero represents the evolving trends of the series news. If it has a positive value, it means that the event abruptly happens, or the event is disappearing.

Though our system performance cannot directly be compared with that of Document Understanding Conference (DUC), it still has the similar performance trend. Our CROUGE-2 score is higher than CROUGE-4 score, and it is reasonable. Because of the limitation of space, our approach 's CROUGE-2 score wasn't listed here.

CRate	Micro	Macro+Micro
10%	19.11%	20.46%
20%	27.56%	29.13%

Table 3: Performance of content selection from macro and micro view with CROUGE-4

5 Conclusion and future work

This paper tries to explore the optimization content selection model for temporal multi-document summarization from different levels of topic detail. We mine the temporal characteristics of articles count, sentences count and events count with a topic changing over time, and proposed a cascaded regression analysis based macro and micro importance discriminative model to guide the content selection. This model not only reduces the complexity of the problem, but also could fully use the temporal characteristics from different levels of topic detail. However, since there are no public evaluation corpus and metrics for temporal

multi-document summarization, our approach cannot compare with others.

In the future, we will explore the more effective model from information fusion view. Considering series news report has the strong temporal characteristic, we will further use the techniques of temporal text mining and temporal information extraction to improve the system performance. We also hope to do some contributions for Chinese temporal multi-document summarization evaluation.

References

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

[2] A. Jatowt and M. Ishizuka. Temporal Web Page Summarization. 5th International Conference On Web Information Systems Engineering, Brisbane, Australia, November 22-24, 2004.

[3] C. Kuan-Yu, L. LUESUKPRASERT, and T. Sengcho. Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling. IEEE Transactions on Knowledge and Data Engineering, pages 1016-1025, 2007.

[4] M. L. Q. W. K. Li, W.J.and Wu. Integrating temporal distribution information into event-based summarization. International Journal of Computer Processing of Oriental Languages, 19:201-222, 2006.

[5] J. Lim, I. Kang, J. J.Bae, and J. Lee. Sentence extraction using time features in multi-document summarization. In Proceedings of the Asia Information Retrieval Symposium, pages 82-932, 2004.

[6] C. Lin. ROU2GE: A Package for Automatic Evaluation of Summaries. Proceedings of the Workshop on Text Summarization Branches Out, pages 25-26, 2004.

[7] I. Mani. Recent Developments in Temporal Information Extraction. Nicolov, N., and Mitkov, R. Proceedings of RANLP, 3, 2004.

[8] R. Swan and D. Jensen. Constructing Topic-Specific Timelines with Statistical Models of Word Usage. Proceedings of the 6th ACM Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 73-80, 2000.