Event Timeline Summarization Using Aging Theory and Incremental Latent Semantic Analysis

Jie Chen*, Ursula Barth, Ingrid Haas, Frank Holzwarth,
Anna Kramer, Leonie Kunz, Christine Reiß,
Nicole Sator, Erika Siebert-Cole, and Peter Straßer

No Institute Given

Abstract. *To reduce peoples time of obtaining the valuable information from amounts of web news, many summarization approaches have been researched. However, most of these approaches ignore the temporal characteristics of news reporting a same event. In this paper, we focus on summarizing the development and changes of events by taking timeline and semantic into consideration. There are three steps in our method. First, we extract hot terms from news which report the same event according to their energy during the time interval we choose (i.e. one day). Second, we use an incremental latent semantic analysis model to recognize the semantic units of news. Third, we construct a semantic text relationship map and choose sentences both important and diverse to generate timeline summaries. Experiment results show that our method can improve the timeline summarization significantly.*

Keywords: Summarization, timeline, aging theory, incremental LSA

1 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 [1] 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, most of these papers 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

* Please note that the LNCS Editorial assumes that all authors have used the western naming convention, with given names preceding surnames. This determines the structure of the names in the running heads and the author index.

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 [2] 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 been 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) [3] 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 [3] 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.

2 Related works

It has been more than 50 years since Luhn [1] proposed automatic summarization. During these years numerous papers have been published on this topic and it has been adopted in many fields. In the beginning, researchers focus on single document summarization. With the rapid development of Internet, the amount of information is increasing in an exponential manner and people easily get lost when faced with such overwhelming information. In order to resolve this kind of

information overload, multi-document summarization has attracted researchers' eyes. Centroid-based method is one of the most popular multi-document summarization methods generating summaries using centroids, position and first-sentence overlapping [9]. While in clustering approaches researchers cluster similar sentences together and select one representative sentence from each main cluster [6, 10]. Later, machine learning [11] and graph-based method [7, 8] are exploited to multi-document summarization. But they all did not consider temporal characteristic. Swan et al. [16] presented temporal characteristics while displaying events under the task of Topic Detection and Tracking. Allan et al. [17] built a temporal summary of news stories to help people monitor changes in news coverage over time. Chieu et al. [18] generated a timeline summary combing a series of events related to an entity with interest and burstiness. Yan et al. [19] built a query-based evolutionary timeline summarization via a balanced optimization framework considering four attributes. However, they all missed the daily hot points of events. Aging theory was put forward by Chen et al. [3] in the task of hot topic detecting, and now we transfer it to our timeline summarization to solve this problem.

In order to improve the qualities of summaries, Brunn et al. [20] proposed to utilize lexical chain to identify the semantic units. Concepts were also exploited to summarization task by [21]. But they all had a heavy reliance on WordNet [22]. Yeh et al. [23] use latent semantic analysis to document summarization without the dependency on lexical resources. Inspired by Yeh, we propose an incremental LSA model for our timeline summarization.

3 Our Approach

3.1 Hot terms identification using Aging Theory

Aging theory is a technique used for tracking life cycles of events, for it consider that every event has a life form with stages of birth, growth, decay and death [2]. Since terms or words are the basic elements of any news report, changes in the content of reports will be reflected by variations in the usage of terms [24]. When a new change of an event occurred, several pieces of news will report it using some special words that can reflect the new change, so frequencies of these words will increase greatly, and with the changes popularity wanes, the frequencies will decrease accordingly. So terms have a similar life span to events, and we can use aging theory to track terms to determine what life stage they are in. Further, we can find out the daily hot spots to improve our summarization. As aging theory using the concept of energy to indicate the liveliness of an event in its life span, our first step is to calculate the energy of terms. The frequency of a word will change as event going on, so we use the association between word and time interval to indicate its energy which is defined as follows:

$$E_{w,t} = F(F^{-1}E_{w,t-1} + \alpha \cdot \chi^2_{w,t}) \tag{1}$$

where $E_{w,t}$ is the energy of word w in time interval$t$, and $E_{w,t-1}$ is the energy of word w in time interval $t-1$, α is the transfer factor, and $\chi^2_{w,t}$ is the contribution

degree of word at the time interval t, which can be computed as presented in [16].

However, no words descripting a special event point will retain popular forever, they will decay over time. In order to represent the word's life span realistically, we cut down the energy of word by a decay factor at the end of every time interval. And if the decayed energy value became negative, we change it to 0.

According to the description above, if the energies of some words increase greatly, we can draw a conclusion that there is a hot event spot. So we need to calculate the variance of word energy next. Here we use standard deviation:

4 Paper Preparation

Springer provides you with a complete integrated LaTeX document class (`llncs.cls`) for multi-author books such as those in the LNCS series. Papers not complying with the LNCS style will be reformatted. This can lead to an increase in the overall number of pages. We would therefore urge you not to squash your paper.

Please always cancel any superfluous definitions that are not actually used in your text. If you do not, these may conflict with the definitions of the macro package, causing changes in the structure of the text and leading to numerous mistakes in the proofs.

If you wonder what LaTeX is and where it can be obtained, see the "*LaTeX project site*" (`http://www.latex-project.org`) and especially the webpage "*How to get it*" (`http://www.latex-project.org/ftp.html`) respectively.

When you use LaTeX together with our document class file, `llncs.cls`, your text is typeset automatically in Computer Modern Roman (CM) fonts. Please do *not* change the preset fonts. If you have to use fonts other than the preset fonts, kindly submit these with your files.

Please use the commands `\label` and `\ref` for cross-references and the commands `\bibitem` and `\cite` for references to the bibliography, to enable us to create hyperlinks at these places.

For preparing your figures electronically and integrating them into your source file we recommend using the standard LaTeX `graphics` or `graphicx` package. These provide the `\includegraphics` command. In general, please refrain from using the `\special` command.

Remember to submit any further style files and fonts you have used together with your source files.

Headings. Headings should be capitalized (i.e., nouns, verbs, and all other words except articles, prepositions, and conjunctions should be set with an initial capital) and should, with the exception of the title, be aligned to the left. Words joined by a hyphen are subject to a special rule. If the first word can stand alone, the second word should be capitalized.

Here are some examples of headings: "Criteria to Disprove Context-Freeness of Collage Language", "On Correcting the Intrusion of Tracing Non-deterministic

Programs by Software", "A User-Friendly and Extendable Data Distribution System", "Multi-flip Networks: Parallelizing GenSAT", "Self-determinations of Man".

Lemmas, Propositions, and Theorems. The numbers accorded to lemmas, propositions, and theorems, etc. should appear in consecutive order, starting with Lemma 1, and not, for example, with Lemma 11.

4.1 Figures

For LaTeX users, we recommend using the *graphics* or *graphicx* package and the `\includegraphics` command.

Please check that the lines in line drawings are not interrupted and are of a constant width. Grids and details within the figures must be clearly legible and may not be written one on top of the other. Line drawings should have a resolution of at least 800 dpi (preferably 1200 dpi). The lettering in figures should have a height of 2 mm (10-point type). Figures should be numbered and should have a caption which should always be positioned *under* the figures, in contrast to the caption belonging to a table, which should always appear *above* the table; this is simply achieved as matter of sequence in your source.

Please center the figures or your tabular material by using the `\centering` declaration. Short captions are centered by default between the margins and typeset in 9-point type (Fig. 1 shows an example). The distance between text and figure is preset to be about 8 mm, the distance between figure and caption about 6 mm.

To ensure that the reproduction of your illustrations is of a reasonable quality, we advise against the use of shading. The contrast should be as pronounced as possible.

If screenshots are necessary, please make sure that you are happy with the print quality before you send the files.

Fig. 1. One kernel at x_s (*dotted kernel*) or two kernels at x_i and x_j (*left and right*) lead to the same summed estimate at x_s. This shows a figure consisting of different types of lines. Elements of the figure described in the caption should be set in italics, in parentheses, as shown in this sample caption.

Please define figures (and tables) as floating objects. Please avoid using optional location parameters like "`[h]`" for "here".

Remark 1. In the printed volumes, illustrations are generally black and white (halftones), and only in exceptional cases, and if the author is prepared to cover the extra cost for color reproduction, are colored pictures accepted. Colored pictures are welcome in the electronic version free of charge. If you send colored

figures that are to be printed in black and white, please make sure that they really are legible in black and white. Some colors as well as the contrast of converted colors show up very poorly when printed in black and white.

4.2 Formulas

Displayed equations or formulas are centered and set on a separate line (with an extra line or halfline space above and below). Displayed expressions should be numbered for reference. The numbers should be consecutive within each section or within the contribution, with numbers enclosed in parentheses and set on the right margin – which is the default if you use the *equation* environment, e.g.,

$$\psi(u) = \int_o^T \left[\frac{1}{2}\left(\Lambda_o^{-1}u, u\right) + N^*(-u)\right] dt\ . \tag{2}$$

Equations should be punctuated in the same way as ordinary text but with a small space before the end punctuation mark.

4.3 Footnotes

The superscript numeral used to refer to a footnote appears in the text either directly after the word to be discussed or – in relation to a phrase or a sentence – following the punctuation sign (comma, semicolon, or period). Footnotes should appear at the bottom of the normal text area, with a line of about 2 cm set immediately above them.[1]

4.4 Program Code

Program listings or program commands in the text are normally set in typewriter font, e.g., CMTT10 or Courier.

Example of a Computer Program

```
program Inflation (Output)
  {Assuming annual inflation rates of 7%, 8%, and 10%,...
   years};
   const
     MaxYears = 10;
   var
     Year: 0..MaxYears;
     Factor1, Factor2, Factor3: Real;
   begin
     Year := 0;
     Factor1 := 1.0; Factor2 := 1.0; Factor3 := 1.0;
```

[1] The footnote numeral is set flush left and the text follows with the usual word spacing.

```
    WriteLn('Year  7% 8% 10%'); WriteLn;
    repeat
      Year := Year + 1;
      Factor1 := Factor1 * 1.07;
      Factor2 := Factor2 * 1.08;
      Factor3 := Factor3 * 1.10;
      WriteLn(Year:5,Factor1:7:3,Factor2:7:3,Factor3:7:3)
    until Year = MaxYears
end.
```

(Example from Jensen K., Wirth N. (1991) Pascal user manual and report. Springer, New York)

4.5 Citations

For citations in the text please use square brackets and consecutive numbers: [1], [2], [4] – provided automatically by LaTeX's `\cite` ... `\bibitem` mechanism.

4.6 Page Numbering and Running Heads

There is no need to include page numbers. If your paper title is too long to serve as a running head, it will be shortened. Your suggestion as to how to shorten it would be most welcome.

5 LNCS Online

The online version of the volume will be available in LNCS Online. Members of institutes subscribing to the Lecture Notes in Computer Science series have access to all the pdfs of all the online publications. Non-subscribers can only read as far as the abstracts. If they try to go beyond this point, they are automatically asked, whether they would like to order the pdf, and are given instructions as to how to do so.

Please note that, if your email address is given in your paper, it will also be included in the meta data of the online version.

6 BibTeX Entries

The correct BibTeX entries for the Lecture Notes in Computer Science volumes can be found at the following Website shortly after the publication of the book:
`http://www.informatik.uni-trier.de/~ley/db/journals/lncs.html`

Acknowledgments. The heading should be treated as a subsubsection heading and should not be assigned a number.

7 The References Section

In order to permit cross referencing within LNCS-Online, and eventually between different publishers and their online databases, LNCS will, from now on, be standardizing the format of the references. This new feature will increase the visibility of publications and facilitate academic research considerably. Please base your references on the examples below. References that don't adhere to this style will be reformatted by Springer. You should therefore check your references thoroughly when you receive the final pdf of your paper. The reference section must be complete. You may not omit references. Instructions as to where to find a fuller version of the references are not permissible.

We only accept references written using the latin alphabet. If the title of the book you are referring to is in Russian or Chinese, then please write (in Russian) or (in Chinese) at the end of the transcript or translation of the title.

The following section shows a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4] and [5], as well as a URL [6]. Please note that proceedings published in LNCS are not cited with their full titles, but with their acronyms!

References

1. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. J. Mol. Biol. 147, 195–197 (1981)
2. May, P., Ehrlich, H.C., Steinke, T.: ZIB Structure Prediction Pipeline: Composing a Complex Biological Workflow through Web Services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148–1158. Springer, Heidelberg (2006)
3. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
4. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181–184. IEEE Press, New York (2001)
5. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)
6. National Center for Biotechnology Information, `http://www.ncbi.nlm.nih.gov`

Appendix: Springer-Author Discount

LNCS authors are entitled to a 33.3% discount off all Springer publications. Before placing an order, the author should send an email, giving full details of his or her Springer publication, to `orders-HD-individuals@springer.com` to obtain a so-called token. This token is a number, which must be entered when placing an order via the Internet, in order to obtain the discount.

8 Checklist of Items to be Sent to Volume Editors

Here is a checklist of everything the volume editor requires from you:

- ☐ The final LaTeX source files
- ☐ A final PDF file
- ☐ A copyright form, signed by one author on behalf of all of the authors of the paper.
- ☐ A readme giving the name and email address of the corresponding author.