

PageRank アルゴリズムを用いた複数時系列文書要約手法への取り組み

柏井香里[†]

[†] お茶の水女子大学理学部情報科学科

小林一郎[‡]

[‡] お茶の水女子大学 基幹研究院自然科学系

1 はじめに

ニュースや新聞記事といった時系列文書は時々刻々と新しい情報が追加されていく。そのような文書の全てを読んで理解することは膨大な時間がかかってしまい現実的でない。複数の情報源からの文書を要約し、時間の経過とともにその内容を把握できる要約手法が望まれる。本研究ではそのことを踏まえて、複数の新聞社による長期にわたる記事から、新しく追加された情報に重きを置いた要約文を時系列順に生成する手法を提案する。

2 時系列複数文書の要約

2.1 先行研究

時系列文書を対象とした要約として、Allan らは temporal summarization を定義した [1]。近年では、Yan ら citeyan2011a により文のランキングアルゴリズムをベースとしたグラフの拡張を行い、異なる時間から1つの平面に文章を射影することによって要約を生成する手法が提案された。LexRank は、Erkan ら [3] によって提案された PageRank [2] に基づいた複数文書要約手法である。この手法では、対象文書中の各文をノードとし、ノードをつなぐエッジを文同士の類似性としてグラフを生成する。多くの文と類似している文は重要度が高いという概念のもと、グラフにおける固有ベクトルの中心性の概念に基づいて文の重要度を計算している。Erkan らは、グラフを生成する際に、類似度の値からエッジの重みを利用する重み付きグラフと、閾値を用いて枝刈りを行う重みなしグラフを提案している。

2.2 提案手法

本研究では、上述した時系列文書要約とグラフを用いた文書要約のそれぞれの手法を踏まえた時系列複数文書要約手法を提案する。提案手法の概要を図1に示す。以下は3日目までの様子である。複数の新聞社による記事を入力とし、各日毎の要約文を出力する。

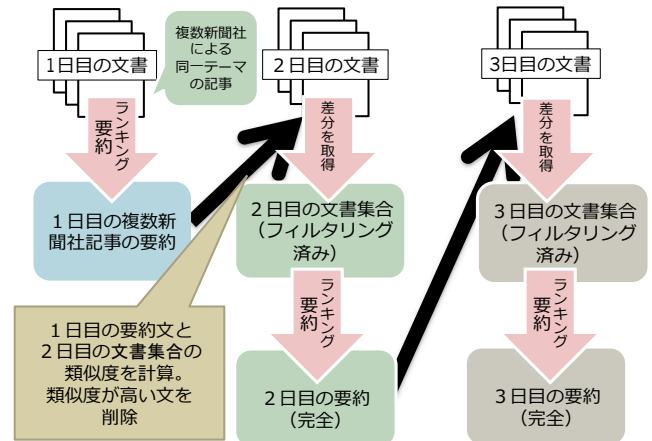


図1: 提案手法の概要

2.3 要約の流れ

本研究では、各文の重要度を決定するためにグラフ構造を用いる。まず、文書集合 $D_t \in \mathcal{D}$ について考える。 t は時刻単位を表し、 $t = \{1, \dots, T\}$ である。ここで、 D_t は時刻 t に属する文書集合を表す。本研究では、時間が経過するとともに新しく文書が追加されることを想定する。Algorithm1 に要約を生成する手順を示す。

入力として、 $\mathcal{D}$, $\mathcal{S}$, ϵ , α を与える。ここで、 $\mathcal{S}$ は出力する要約の候補となる文集合、 α は前日の要約文と当日の文との類似度の閾値であり、 ϵ は要約として出力する文の数である。文集合 $\mathcal{S}_t$ に含まれる文で構成されるグラフを考える。文のランキングアルゴリズムに [3] で提案される LexRank アルゴリズムを用いた。

3 実験

3.1 実験設定

使用したデータ、正解データなど実験に関する設定を記載する。対象データには、Tran らが提供しているタイムライン要約のためのデータセットを用いる。このデータセットは以下の論文で使用されている [5][6]。これらは、複数のニュース源から集められた9つのトピックに属している新聞記事である。表1に用いたデータセットの詳細を示す。

An Approach to Multiple Time-series Documents Summarization using PageRank Algorithm

[†]Kaori KASHIWAI(g1220515@is.ocha.ac.jp),

[‡]Ichiro KOBAYASHI(koba@is.ocha.ac.jp)

表 2:生成された時系列の要約文書

2010-04-24	The Bureau of Ocean Energy Management , Regulation and Enforcement -LRB- BOEMRE -RRB- , formerly known as the ...
2010-04-25	Neither the Coast Guard nor their employers have released their names , though several of their families have come forward...
2010-04-26	It is a big change from yesterday ... This is a very serious spill , absolutely ,” said Rear Adm Mary Landry ...
2010-04-27	“ Eventually , things will return to normal . ” Forecasters say it could wash ashore within days near delicate wetlands ...

Algorithm 1 要約のプロセス

```

Input:  $D$ ,  $S$ ,  $\epsilon$ ,  $\alpha$ ,  $l$ 
 $S = \{ \}$ 
 $\epsilon \leftarrow \text{threshold1}$ 
 $\alpha \leftarrow \text{threshold2}$ 
for  $t = 0$  to  $T$  do
  if  $t=0$  then
     $S_t \leftarrow D_t$ 
  else
     $S_t = [ ]$ 
    for  $d$  to  $|D_t|$  do
      for  $s$  to  $|S_{t-1}|$  do
        if  $\text{similarity}(d, s) < \alpha$  then
           $S_t \leftarrow d$ 
        end if
      end for
    end for
    ranking  $S_t$  with LexRank
    if  $\text{length of } S_t > \epsilon$  then
       $S'_t \leftarrow \text{top } \epsilon \text{ sentences of } S_t$ 
    else
       $S'_t \leftarrow S_t$ 
    end if
  end if
   $S \leftarrow S'_t$ 
end for
return  $S$ 

```

3.2 実験結果

トピック BP Oil Spill について生成した要約文の一例を表 2 に示す。

3.3 考察

BP Oil Spill について、約 1400 文書を各日 10 文までの合計 1700 文程度と、総文数と比べてかなり短く要約をすることができた。入力された文書の各文と前日の要約文との類似度を計算し、前日の要約で既に登場した情報を含む文を取り除くことによって、出力された要約文に冗長性はあまり見られず、新しく追加された情報を把握しやすくなった。また、複数の新聞社の記事を要約しているため、要約文はどの新聞社にも載っている重要で信憑性の高いものになった。

表 1:ニュース資源

トピック	ニュース源	文書数	正解の文数
BP Oil Spill	BBC	293	98
BP Oil Spill	Foxnews	286	52
BP Oil Spill	Guardian	288	307
BP Oil Spill	Reuters	298	30
BP Oil Spill	Washingtonpost	296	19

4 おわりに

LexRank による重要文抽出と、前日の要約との冗長性を避ける文抽出により、各日毎の重要となる情報を含む文から要約文を生成することができた。これにより時系列に沿った要約文生成を行った。しかし、閾値の設定や出力する文の数の上限などを、要約文がより見やすくなるように設定し直す必要がある。また、今回正解データに基づいた評価を行っていないので、今後、正解データに基づき ROUGE を用いて評価を行う。

参考文献

- [1] James Allan, Rahul Gupta, and Vikas Khandelwal, Temporal Summaries of News Topics, In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001.
- [2] Sergey Brin and Lawrence Page, The Anatomy of Large-scale Hypertextual Web Search Engine, Computer Networks and ISDN Systems, pp. 107-117, 1998
- [3] Gunes Erkan and Dragomir R. Radev, LexRank: Graph-based Lexical Centrality as Saliency in Text Summarization, Journal of Artificial Intelligence Research, pp. 457-479, 2003.
- [4] C. Lin, ROUGE: a Package for Automatic Evaluation of Summaries, In Proceedings of the Workshop on Text Summarization Branches Out, pp. 74-81, 2004.
- [5] G. B. Tran, Tuan A. Tran, N. Tran, M. Alrifai, and N. Kanhabua, Leveraging Learning To Rank in an Optimization Framework for Timeline Summarization, SIGIR, 2013.
- [6] G. B. Tran, M. Alrifai, and D. Q. Nguyen, Predicting Relevant News Events for Timeline Summaries, In Proceedings of the 22nd international conference on World Wide Web Companion, pages 91-92. International World Wide Web Conferences Steering Committee, 2013.
- [7] R. Yan, L. Kong, C. Huang, X. Wan, X. Li, and Y. 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, 2011a.