# 時系列文書を対象としたグラフに基づく文書要約への取り組み

鈴木　聡子　　小林 一郎
お茶の水女子大学大学院人間文化創成科学研究科理学専攻
{suzuki.satoko, koba}@is.ocha.ac.jp

## 1　はじめに

情報技術の発展に伴い，我々は大量のデータの蓄積・閲覧が可能となった．そのため，重要度の高い情報や利用者が欲している情報が選択されやすくなるための情報検索や，膨大な情報の中から効率良く内容を把握するための自動要約において，より高精度な技術の必要性が高まっている．自動要約の研究は，新聞記事や学術論文からブログや twitter まで，対象とされる文書は様々である．ここで新聞記事に着目すると，短期的な話題と長期的な話題が存在する．長期的に記載された話題の場合，話題の概要に加えて，時間に伴って内容がどのように変化したかを知りたいという読み手の欲求が考えられるが，通常の複数文書要約タスクにおいては，この欲求を満たすことは難しい．また長期的な話題に関する記事では，その話題に関して新たに出来事が起こった場合，記事が追加される．そのため，文書群の更新とともに，要約の内容も更新される必要がある．

本研究では，グラフ構造を用いたランキングアルゴリズムを拡張し，時系列文書を対象とし話題変遷の把握が可能な要約の生成を目的とする．

## 2　関連研究

### 2.1　複数文書要約

文書要約は，手法で大別すると圧縮型と抽出型に分けられる．抽出型の要約手法は，重要性に関するスコアを単位（文やパラグラフ）ごとに割り当て，スコアが高い順に抽出する手法であり，代表的な手法として重心法に基づく手法 [8] や組み合わせ最適化問題に帰着させた手法，グラフのランキングアルゴリズムに基づいた手法 [3][6][13] などが挙げられる．

### 2.2　時系列文書を対象とした要約

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]．最近では，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 [11] や，関連性・被覆率・結合性・多様性のような異なる側面の組み合わせを考慮した関数の最適化により要約を生成する手法 [12] が Yan らにより提案された．また Jiewi らは，トピックの進化パターンを考慮するために Evolutionary Hierarchical Dirichlet Process(EHDP) と呼ばれる新しいモデルの提案を行った [4].

### 2.3　グラフに基づいた文書要約

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．

## 3　提案手法

図 1 に提案手法のイメージ図を示す.

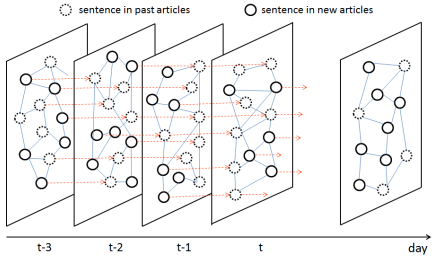


図 1: 提案手法のイメージ図

### 3.1 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 $D_t \in \boldsymbol{D}$ について考える．$t$ は時刻単位を表し，$t$={ 1,⋯,$T$ } 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.

**Algorithm 1** 要約のプロセス

**Input:** $\boldsymbol{D}$, $\boldsymbol{S}$, $\epsilon$, $l$
$\boldsymbol{S}$ = { }
$\epsilon \leftarrow$ threshold
**for** $t$ = 0 to $T$ **do**
  $\boldsymbol{S}' \leftarrow \boldsymbol{S}$ + $D_t$
  ranking $\boldsymbol{S}'$ with LexRank
  **if** length of $\boldsymbol{S}' > \epsilon$ **then**
    $\boldsymbol{S} \leftarrow$ top $\epsilon$ sentences of $\boldsymbol{S}'$
  **else**
    $\boldsymbol{S} \leftarrow \boldsymbol{S}'$
  **end if**
**end for**
**return** top $l$ sentences of $\boldsymbol{S}$

入力として，$\boldsymbol{D}$，$\boldsymbol{S}$，$\epsilon$，$l$ を与える．ここで，$\boldsymbol{S}$ は出力する要約の候補となる文集合，$\epsilon$ は閾値であり，$l$ は要約として出力する文の数である．文集合 $\boldsymbol{S}'$ 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．本研究では，閾値による枝刈りを行わない重みなしグラフを適用した.

### 3.2 グラフサイズの設定

提案手法では，グラフの大きさを制限する．グラフの大きさが設定した閾値を超えた場合には，閾値以下の大きさにグラフを縮小する．常に閾値以下のグラフサイズを保った状態で文のランキングを行い，要約が必要なタイミングでスコアの高い文から要約の候補として抽出する.

## 4 実験

### 4.1 データ

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 [9][10]．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.

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 |
| H1N1 Influenza | BBC | 122 | 40 |
| H1N1 Influenza | Guardian | 76 | 34 |
| H1N1 Influenza | Reuters | 207 | 23 |
| Haiti Earthquake | BBC | 296 | 86 |
| Iraq War | Guardian | 344 | 410 |
| Libya War | CNN | 398 | 81 |
| Syrian Crisis | BBC | 308 | 31 |

### 4.2 実験設定

比較のためのベースラインとして，ランダムに抽出したものと重みなしグラフによる LexRank を用意する．全てのシステムにおいて，生成する要約の長さは，各トピックにおける正解要約の文の長さと等しいものとした．また，前処理として ‘a’ や ‘the’ といったありきたりな語であるストップワードの除去と，語尾の異なるものを同一とみなすためのステミング処理を全てのシステムにおいて行った．ステミングには Porter のアルゴリズム [7] を用いる.

### 4.3 評価手法

人手で作成された正解要約と，最終的な時点で出力された要約を比較することでシステムの評価を行う．評価には，ROUGE[5] を用いる．今回は，評価にユニグラムを使用した ROUGE-1 における再現率と F 値を評価に用いる．また，ステミング処理を行った後，stopwords を含める場合と含めない場合で評価を行う．以下，前者を with，後者を without として示す.

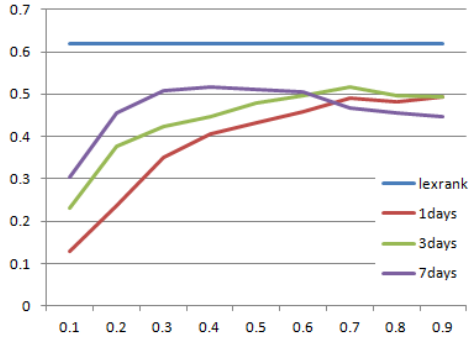


図 2: 再現率/with

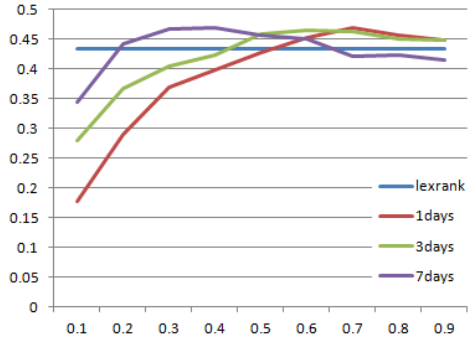


図 3: F 値/with

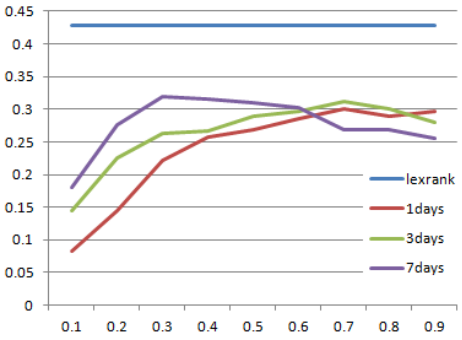


図 4: 再現率/without

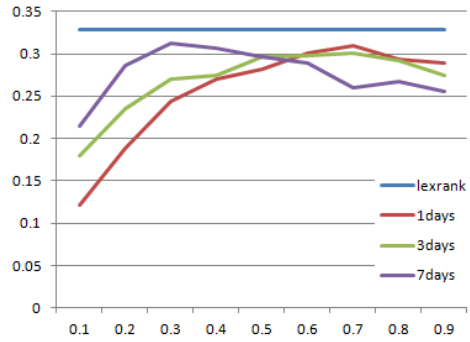


図 5: F 値/without

### 4.4 結果

まず，グラフを固定値とした場合に総文数に対するグラフサイズの割合によりシステムを 2 つに分けたシステム全体の評価結果を表 2 に示す.

表 2: グラフサイズを固定値とした結果

| | with | | without | |
|---|---|---|---|---|
| | 再現率 | F 値 | 再現率 | F 値 |
| random | 0.509 | 0.464 | 0.314 | 0.304 |
| Lexrank | 0.620 | 0.434 | 0.430 | 0.328 |
| 提案手法 A | 0.607 | 0.473 | 0.424 | 0.340 |
| 提案手法 B | **0.658** | **0.490** | **0.495** | **0.377** |

提案手法 A は，グラフサイズが総文数に対して半分程度だった場合の要約を評価したものである．全ての評価値において，最も値の高い数値を太字で記す．それに対し，提案手法 B はグラフサイズが総文数に対して 2/3〜3/4 程度のサイズの要約を評価したものである．全ての評価値において，システムとして最も良い精度を示したのは提案手法 B である．提案手法 A では，どの評価値でも LexRank を下回る結果となった．

次に，グラフサイズを比率とした場合の実験結果を図 2〜図 5 に示す.

この実験では，グラフの更新期間を 1，3，7 日と変更して実験を行った．各グラフの縦軸は評価値を示し，横軸は比率の変化を示している．また比較として，LexRank での実験結果を載せる．グラフより，ストップワードを含む場合の F 値を除き，提案手法は LexRank を下回る結果となっている．また，LexRank と比較してストップワードの有無による精度の差が大きい．更新期間ごとの精度を見てみると，1 日ごとにグラフを更新した場合はどの評価値においても，比率を小さく設定した場合には精度が最も低いが，比率を大きくするほど精度は上がり，0.7〜0.9 では他の更新期間と比較しても良い精度を示している．3 日ごとにグラフを更新した場合，他の更新期間と同様，比率を大きくするほど精度が良くなっており，1 日ごとの更新と比較すると，比率が小さい場合の精度が良いことが分かる．最後に，7 日ごとにグラフを更新した場合では，比率が小さい場合の精度は他の 2 つと比べると最も良い結果を示しているが，0.4 辺りから徐々に精度が下がり，0.7 以上ではどの評価値においても完全に他の手法を下回っている．

## 5 考察

グラフサイズを固定値に設定した実験では，総文数に対するグラフのノード数の比率でシステムを分けて評価を行ったところ，2/3〜3/4 程度のグラフサイズに設定した場合に最も良い精度を示した．このことから，全ての文を用いてグラフに基づくランキングにより要約を生成するよりも，重要度の低い文を取り除いた状態でランキングを行った方が精度が上がることが確認できた．しかし，対象としたい時系列データでは，本来ならば記事がどこまで増加するかは分からないため，固定値としてグラフサイズを決定することは実際問題としては，困難であると思われる．

グラフサイズを比率として制限していく方法は，日々増加し続けるデータへの適用として，ある程度有用であると考えられるが，実験結果では良い精度を得ることが出来なかった．この結果から，グラフを更新していく過程において，グラフのノード数が小さい場合，ランキングの精度が悪く，初期段階でのノードの消去が悪影響の要因であると考えられる．

次に，グラフの更新期間の設定に関して，更新期間が短い程，グラフの比率が小さい場合には要約の精度が悪くなる．そのため，更新の際には多くの文を維持しておく必要がある．一方，更新期間が長い程，グラフの比率が小さい場合の精度が高いが，比率を大きくした場合に精度の低下が見られる．そのため，現段階では最適なパラメータを断定することは困難である．

## 6 おわりに

本研究では，時系列文書を対象とした要約生成に向けて，グラフに基づく要約手法の提案を行った．また，提案手法に従って実験及び評価を行った．結果として，時間の経過に伴いグラフの更新を行う際に，重要度の低い文を消去することによって，要約の精度が上がったことが確認できた．しかしグラフサイズの設定方法が困難であり，今後検討する必要があると考える．また，現段階では最終時点で生成された要約のみを評価に用いているが，様々な時点における生成要約の評価実験によって，提案手法の評価を行っていくことを今後の課題とする．

## 謝辞

本研究の一部は，公益財団法人の栢森情報科学振興財団からの支援により達成されました．ここに感謝の意を表します．

## 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 Anatony 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 Salience in Text Summarization，Journal of Artificial Intelligence Research, pp. 457-479，2003.

[4] J. Li and S. Li，Evolutionary hierachical dirichlet process for timeline summarization，In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL'13, pages 556-560. Association for Computational Linguistics，2013.

[5] C. Lin，ROUGE: a Package for Automatic Evaluation of Summaries，In Proceedings of the Workshop on Text Summarization Branches Out, pp. 74-81，2004.

[6] R. Mihalcea and P. Tarau，TextRank: Bringing order into texts，In Proceedings of EMNLP-03 and the 2004 Conference on Empirical Methods in Natural Language Proceeding，2004.

[7] M.F. Porter，An algorithm for suffix Stripping，Program, Vol. 14 No.3,pp.130-137，1980.

[8] D. R. Radev, H. Jing, and M. Budzikowska，Centroid-based Summarization of Multiple Documents: Sentence Extraction，Utility-based Evaluation，and User Studies, ANLP/NAACL Workshop on Summarization，2000.

[9] 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.

[10] 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.

[11] 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.

[12] R. Yan, C. Huang, X. Wan, J. Otterbacher, X. Li, and Y. Zhang，Timeline Generation Evolutionary Trans-Temporal Summarization，In Proceedings of the Conference on Empirical Method in Natural Language Processing，2011b.

[13] R. Yan, X. Wan, Y. Zhang, and X. Li，Herachical Graph Summarization: Leveraging Hybrid Information through Visible and Invisible Linkage，PAKDD，2012.