1P3-2

オンライン議論掲示板における 多層グラフを用いた自動要約のための重要文抽出手法の提案

Extracting Sentences for Automated Summarizing using Multi-layer Graph on Online Discussion Forum

北川涼太 *1　藤田桂英 *1
Ryota Kitagawa　Katsuhide Fujita

*1東京農工大学 工学府 情報工学専攻
Department of Computer and Information Sciences, Graduate School of Engineering, Tokyo University of Agriculture and Technology

On online discussion forums with the large number of participants, the facilitator is necessary to promote discussions smoothly following the discussion theme and reach final agreements. An automated summarization is one of the effective tasks to support the facilitator. In this paper, we propose the method of extracting sentences from a set of posts in a target thread for automated summarization. We define a multi-layer graph consisting "user", "post", "sentence" and "word" layers for online discussion forum. In addition, the scores computed in each layer based on eigenvector centrality are aggregated into a sentence layer. After that, sentences with high scores are extracted until length of summary exceeds given length. As a result of the survey questionnaire, the proposed method outperformed the baseline modified the previous work in terms of comprehension and focusing.

1. はじめに

近年，様々な局面でオンラインディスカッションが活用される機会が増え，時間的・空間的に離れたユーザであっても議論を交わすことを可能とした新しいオンライン議論プラットフォームが開発されている．その一例として，*Collagree*[伊藤 15] は大規模な人数のユーザが参加するタウンミーティングや社内会議といった場での利用を目的とし，議論を発散・収束・集約というフェーズに切り分け，各フェーズで効果的に議論を支援するための機能を提供している．しかし，参加人数の増加や実施期間の長期化など議論の大規模化に伴い意見の投稿数も増加するため，ユーザやファシリテータが議論の内容を把握するのが困難という問題がある．例えば，ユーザが意見を投稿する際，既にほかのユーザが投稿した内容と重複していないか確認するために，現時点までになされた議論にすべて目を通すことは現実的ではない．以上の問題を解消するために，議論掲示板における自動要約は重要なツールとなりうる．

Collagree に代表される一般的な議論掲示板では，はじめにユーザが話し合う議題が与えられ，ファシリテータの進行に従いながら議論を進めていくことが多い．議論掲示板の特徴として，一つの話題・論点に関する一連の投稿意見集合によって構成されるスレッド構造が挙げられる．図 1 に示すように，スレッドに含まれる投稿間には必ず返信関係が成り立っており，返信関係にある二つの投稿は同じ話題について言及している可能性が高い．羽鳥ら [羽鳥 10] は，これらの投稿間の話題的な繋がりと関連語の連鎖を表した語彙的連鎖に着目し拡張した PageRank[Page 99] を用いて，各投稿の重要文およびスレッドのトピックの抽出手法を提案した．

議論掲示板に含まれる投稿意見からなるテキストデータを対象とした要約では，文そのものの重要度に返信関係などの議論掲示板特有の非テキスト情報を加え，結論に至るまでの過程として議論の流れを反映させた要約文を生成するのが望ましい．返信関係は投稿意見集合のネットワークとして表現できるため，グラフベースのアプローチを用いるが有効である．また，その他のテキストの表層情報などについても対応するグラフを定義し，すべての層を積み重ねた多層グラフを構築することによって重要文抽出を行うことで重要文抽出の性能が向上する可能性がある．多層グラフを用いた複数文書要約のアプローチとして，渋木ら [渋木 14] の Web 文書を対象としたグラフベースのアプローチがある．文書層・文層・単語層の 3 層から構成される多層グラフで対象とする文書集合を表現し，各層で求めたノードの重要度を文層に集約させることで重要文を抽出する．

本論文では，オンライン議論掲示板に対応した多層グラフを提案し，スレッドごとの要約を自動生成するための重要文抽出手法を提案する．渋木らは，3 層の多層グラフを定義したが，本論文では投稿間の返信関係に基づいた重要文抽出を行うため

図 1: *Collagree* のスレッドの例

連絡先: 北川涼太，東京農工大学小金井キャンパス，東京都小金井市中町 2 丁目 24 番 16 号，042-388-7141，kitagawa@katfuji.lab.tuat.ac.jp

に，ユーザ層・意見層・文層・単語層の 4 層から構成される多層グラフを提案する．そして，各層において固有ベクトル中心性に基づいて求めたノードの重要度を文層に集約させ，要約長を満たすまで重要度の合計値が大きい文を抽出し，得られた重要文集合を要約として出力する．

また，*Collagree* の議論データに対して，提案手法と既存手法により重要文抽出を行い要約文を生成し，被験者による評価アンケートを実施する．内容理解，焦点，非冗長性，網羅性の４項目で比較する．

以下に本論文の構成を示す．2. では主なグラフベースの複数文書要約の関連研究を示す．3. では返信関係を考慮した自動要約手法について提案する．4. で評価実験結果を示した後，5. で本論文のまとめと今後の展望について述べる．

2. 関連研究

新聞記事や学術論文などの一般的なテキストを対象とした複数文書要約は，原文書中の情報をいかに多く被覆し，かつ冗長性を排除するかに焦点が当てられており，重要と思われる言語的な単位 (文，文節など) を抽出し文意に矛盾の生じない適切な順序に並び替えることによる抽出型のアプローチが多い．グラフベースアルゴリズムを応用した複数文書要約として，Erkan ら [Erkan 04] は PageRank を複数文書要約に拡張した LexRank を提案した．対象とする文書集合から，ノードが文書，エッジが文書間のコサイン類似度に対応したグラフを構築し，PageRank[Page 99] を拡張することによって，隣接するノードの重要度も考慮に入れながら自身のノードの重要度を求める固有ベクトル中心性に基づき重要文抽出し，要約文を生成する．

羽鳥ら [羽鳥 10] は，議論掲示板のスレッドをトピックの観点から俯瞰することを目的とし，各投稿の重要文とスレッドのトピックを抽出するアプローチを提案した．重要文抽出においては，重要文/非重要文に特徴的な手がかり語，返信関係にある投稿，伝搬される語彙的連鎖を優先的に重み付ける項を PageRank に追加することによって，スレッド構造のより詳細なモデリングを行った．

グラフベースアルゴリズムに基づく別の複数文書要約として，Co-HITS-Ranking を用いた Hu ら [Hu 10] らのアプローチがある．Hu らは，(1) クエリや重要な文と重くリンクされた文が重要な文である，(2) 重要な文書に包含される文が重要な文であるという二つの仮説を立て，文書層と文層の 2 層から構成される多層グラフで対象とする文書集合を表現し，異層間の情報を統合することで要約の精度の向上を達成した．

渋木ら [渋木 14] は Hu らの Co-HITS-Ranking を文書層・文層・単語層の 3 層に拡張したアプローチを提案した．文層においては，Bag of Word(BoW) の代わりに Basic Element[Hovy 06] という最小の意味的な単位を，単語層においては，単語の概念を用いたシソーラス距離をノード間の類似度に導入し，多様な観点からの情報を統合した重要文抽出を行った．

3. 多層グラフによる自動要約手法

3.1 前処理

テキストデータの前処理として，URL と空行は重要文抽出を行う上で必要ないと判断し削除する．また，投稿意見の本文中で，“。”，“. ”，“？”，“！” のいずれかが一回以上連続して出現したとき，その記号を区切りとして文分割を行うが，括弧の中の表現である場合は例外として文分割を行わない．形態素解析には MeCab[Kudo 04] を用い，活用のある品詞については終止形に揃える処理を行う．

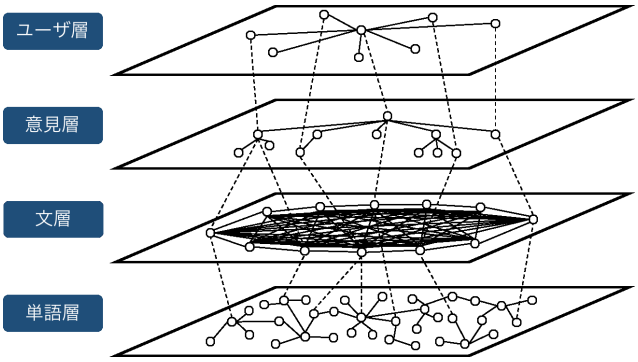

図 2: 多層グラフのイメージ

3.2 多層グラフの構築

本論文では，議論掲示板の一つのスレッドに含まれる複数の投稿意見集合から重要文抽出を行う手法を提案する．渋木ら [渋木 14] の文書層・文層・単語層の 3 層から構成される多層グラフでは，議論掲示板の特徴の一つであるユーザ間の返信関係を捉えられない．以上から，新たにユーザ層を加えた 4 層から構成される多層グラフを定義する．図 2 は，提案手法の多層グラフのイメージを示しており，ユーザ層・意見層・文層・単語層の 4 層を持つ．図 2 の実線は，同じ層でのノード間のエッジを表し，点線は異なる層のノード間の包含関係を表している．

【ユーザ層】

ノードがユーザ，エッジが各ユーザがこれまでに投稿した意見の返信先ユーザを表す．提案手法の要約対象はあくまでスレッドであるが，ユーザ層では掲示板全体を対象とすることで，対象スレッドだけではなく，ほかのスレッドでも大きな発言権を持つユーザの要素を考慮に入れている．

【意見層】

ノードが投稿意見，エッジが投稿意見間の返信関係を表している．ユーザ層とは異なり，対象スレッドに含まれる投稿意見の返信関係のみで意見層を構築する．これにより，多く返信されている意見は重要であるとし，特に話題の導入や問題提起を担っているスレッドの根意見に含まれる文を選出する効果が期待される．

【文層】

ノードが各投稿意見に含まれる文を表し，すべての組み合わせのノード対でエッジを張り，文間のコサイン類似度をその重みとする．これにより，スレッド内部で多くのユーザに言及されているトピックを多く被覆する文を抽出する．

【単語層】

ノードが各文に含まれる単語，エッジが同一の文中における共起の有無を表している．文よりも粒度の細かい単語レベルでのトピックを同定することで，そのスレッド内でキーワードとなっている単語を要約に盛り込むことを目的とする．

3.3 ノードの重要度の計算および伝搬

前節で定義した多層グラフの各層において，ノードの重要度を計算する．ユーザ層・意見層・単語層では，PageRank[Page 99] を用いる．PageRank は，Web ページのハイパーリンクから構成される有向グラフを用いて，Web ページの順位付けを行

うためのアルゴリズムである．重要度の高い Web ページにリンクされている Web ページもまた重要度が高くなる特徴を持つ．PageRank は以下のように定義される．

$$R(n_i) = \frac{1-d}{N} + d \sum_{n_j \in In(n_i)} \frac{R(n_j)}{|Out(n_j)|} \quad (1)$$

ここで，$n_1, n_2, \ldots, n_N$ はノード，$In(n_i)$ は n_i にリンクしているノードの集合，$Out(n_j)$ は n_j がリンクしているノードの集合，N はノードの総数を表す．

一方，文層では PageRank を複数文書要約に拡張した LexRank[Erkan 04] を用いる．LexRank は以下のように定義される．

$$R(n_i) = \frac{1-d}{N} + d \sum_{n_j \in Adj(n_i)} \frac{sim(n_i, n_j)}{\sum_{n_k \in Adj(n_j)} sim(n_k, n_j)} R(n_i) \quad (2)$$

$sim(n_i, n_j)$ は n_i, n_j のコサイン類似度，$Adj(n_i)$ は n_i に隣接するノードの集合を表す．d は式 (1),(2) ともに一定の割合で非隣接ノードのジャンプする制動係数 $[0, 1]$ であり，Page らに倣い値を 0.85 に設定した．

ユーザ層・意見層・文層・単語層のそれぞれでノードの重要度を計算した後，図 2 中の点線で示した包含関係に従って文層にほかの層の重要度を伝搬させ，以下の式で文層のノードの最終的なスコアを算出する．

$$\begin{aligned} Score(n_{\text{sent}}) = {} & R(n_{\text{user}}) + R(n_{\text{post}}) \\ & + R(n_{\text{sent}}) + \sum_{n_{\text{word}} \in n_{\text{sent}}} R(n_{\text{word}}) \end{aligned} \quad (3)$$

$n_{\text{user}}, n_{\text{post}}, n_{\text{sent}}, n_{\text{word}}$ は，それぞれユーザ層・意見層・文層・単語層におけるノードを表している．式 (3) では，文ノードの LexRank 値に，ユーザ層と意見層については 1 対 1 で対応するノードの PageRank 値を伝搬し，単語層についてはその文ノードが包含するすべての単語ノードの PageRank 値を伝搬し加算する．最後に，スコアが上位の文から順に，指定した要約率を満たすまで重要文を選択し，投稿時刻が若い順に並び替えて要約を生成する．

4. 評価実験

4.1 ベースライン

提案手法の評価実験を行う際の比較対象となるベースラインは，渋木ら [渋木 14] の文書層・文層・単語層の 3 層からなる多層グラフを用いた要約手法を参考とした．文書層を意見層に置き換え，各層におけるノード間のエッジは彼らの手法に倣い，以下の類似度によって重み付けしたグラフを用いた．

$$sim_{BoW}(n_1, n_2) = \frac{cbow(n_1, n_2)}{|bow(n_1) \cup bow(n_2)|} \quad (4)$$

$$sim_{BoBE}(n_1, n_2) = \frac{cbobe(n_1, n_2)}{|bobe(n_1) \cup bobe(n_2)|} \quad (5)$$

$$sim_{TD}(n_1, n_2) = \frac{D_{TD} - depth(n_c)}{D_{TD}} \quad (6)$$

ここで，$cbow(n_1, n_2)$ は n_1 のテキストと n_2 のテキストに共通して含まれる単語の異なり数，$bow(n)$ は n のテキストに含まれる単語の集合である．$cbobe(n_1, n_2)$ は n_1 のテキストと n_2 のテキストに共通して含まれる Basic Element の異なり数，$bobe(n)$ は n のテキストに含まれる Basic Element の集合を示す．D_{TD} はシソーラス上のルートノードから葉ノードまでの距離，n_c は n_1 と n_2 の共通上位ノード，$depth(n)$ はルートノードから n までの距離である．

以上の 3 層グラフの各層において，PageRank によりノードの重要度を計算した後に，内包関係に基づき式 (3) に従って文層に重要度を伝搬させる．テキストデータの前処理および抽出した重要文の並び替えは，提案手法と同じである．

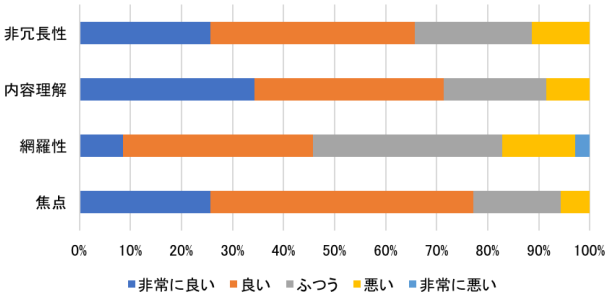

図 3: 提案手法の評価結果

4.2 実験設定

重要文抽出の効果を検証するために，大学院生 5 人に評価アンケートを実施し，提案手法とベースラインの重要文抽出に関して比較実験を行った．対象データは，*Collagree* の過去の議論データから取得した 7 スレッドを対象する，議論データの平均ユーザ数が 8.29，平均投稿意見数が 10.86 であった．評価項目は以下の 4 項目である．

【非冗長性】同じ情報が繰り返されていないか
非常に良い：繰り返されていない ～ 非常に悪い：繰り返されている

【内容理解】要約を読んで原文書の大意を把握できるか
非常に良い：把握できる ～ 非常に悪い：把握できない

【網羅性】原文書の重要と思われる情報が不足していないか
非常に良い：不足していない ～ 非常に悪い：不足している

【焦点】要約全体と関係のない情報が含まれていないか
非常に良い：含まれていない ～ 非常に悪い：含まれている

原文書に対する要約の圧縮率は 30%とした．

4.3 実験結果

図 3,4 に，それぞれ提案手法とベースラインにより抽出した重要文に対する評価アンケートの結果を示す．内容理解と焦点については，“非常に良い” または “良い” と回答した割合が提案手法がベースラインよりも高かった．しかし，非冗長性と網羅性については，ベースラインが提案手法を上回る結果となった．

提案手法では，ユーザ層と意見層で返信関係に基づくグラフを用いて，議論に積極的なユーザや意見に高い重要度を付与することになるので，より多く返信されているスレッドの根意見（最初の投稿）から重要文を多く選択する傾向が強くみられた．スレッドの根意見は，話題の導入や問題提起の役割を担っ

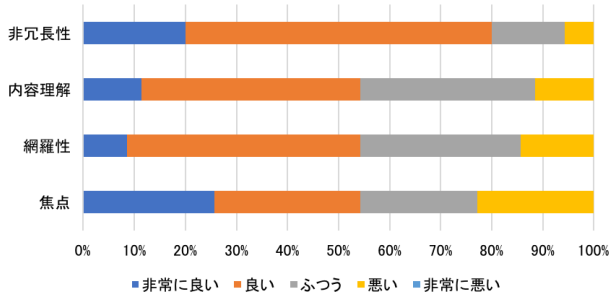

図 4: ベースラインの評価結果

ていることが多いため，スレッドを対象とした要約には根意見に属する文を含めることも重要であり，内容理解の向上に寄与したと思われる．一方，ベースラインは根意見以外の意見からも重要文を抽出しているため，提案手法よりも高い網羅性を示した．また，提案手法の非冗長性が低かった結果から，より少ない意見から重要分を抽出すると類似した表現が繰り返される傾向にあることが明らかになった．

5. おわりに

本論文では，オンライン議論掲示板のスレッドを対象とした自動要約のための多層グラフを用いた重要文抽出手法の提案・評価を行った．議論掲示板が持つスレッド構造を利用し，返信関係により構築したユーザ層・意見層，テキストの表層情報により構築した文層・単語層の 4 層から構成される多層グラフによって対象とする投稿意見集合を表現した．多層グラフの各層において，PageRank および LexRank を用いて求めた固有ベクトル中心性をノードの重要度とし，ユーザ層・意見層・単語層の重要度を包含関係に従って文層に伝搬した後，重要度の合計値が高い文を重要文として抽出した．また，提案手法の効果を検証するために，被験者への評価アンケートを通じてベースラインとの比較を行った．内容理解と焦点に関して，提案手法がベースラインを上回る結果が得られた．

今後の展望として，各層のノードの重要度を文層に伝搬する際に，各層の重要度を考慮した重み付けを行うことが挙げられる．提案手法では，ユーザ層と意見層の影響が強く，スレッドの根意見から多くの重要文を抽出してしまうという問題があった．これらの層の重みを小さくすることで，他の意見からも重要文を抽出することを可能とし，非冗長性と網羅性における精度の向上が見込めると考えられる．

謝辞

本研究は，JST，CREST の支援を受けたものである．

参考文献

[Erkan 04] Erkan, G. and Radev, D. R.: LexRank: Graph-based Lexical Centrality As Salience in Text Summarization, *J. Artif. Int. Res.*, Vol. 22, No. 1, pp. 457–479 (2004)

[Hovy 06] Hovy, E., Lin, C.-Y., Zhou, L., and Fukumoto, J.: Automated summarization evaluation with basic elements, in *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006)*, pp. 604–611Citeseer (2006)

[Hu 10] Hu, P., Ji, D., and Teng, C.: Co-hits-ranking based query-focused multi-document summarization, in *Asia Information Retrieval Symposium*, pp. 121–130Springer (2010)

[Kudo 04] Kudo, T., Yamamoto, K., and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis., in *EMNLP*, Vol. 4, pp. 230–237 (2004)

[Page 99] Page, L., Brin, S., Motwani, R., and Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web., Technical Report 1999-66, Stanford InfoLab (1999)

[伊藤 15] 伊藤 孝行, 奥村 命, 伊藤 孝紀, 秀島 栄三：多人数ワークショップのための意見集約支援システム Collagree の試作と評価実験, 日本経営工学会論文誌, Vol. 66, No. 2, pp. 83–108 (2015)

[羽鳥 10] 羽鳥 潤, 村上 明子：スレッド構造と語彙的連鎖を用いたオンラインディスカッションからの重要文・トピックの抽出, 言語処理学会第 16 回年次大会 (NLPs2010) (2010)

[渋木 14] 渋木 英潔, 森 辰則：クエリ指向の要約のためのクエリ指向の要約のための異種情報を統合したグラフベースの重要文抽出手法の提案, 言語処理学会第 20 回年次大会 (NLPs2014) (2014)