Vol.2015-ICS-181 No.5
2015/12/22

議論掲示板におけるスレッド構造と時系列を考慮した自動要約

北川 涼太[1,a)]　藤田 桂英[1]

概要：Web 上で議論を行うことを目的にした議論掲示板では，議論が深まるにつれて話題が多様化し，全体構造が複雑化する．そのため，ユーザがこれまでに行われた議論の内容を理解するために，議論内容の要約を提示することが有効である．本研究では，一つの話題を扱うスレッドを対象に，時系列や返信関係，ユーザ情報といった要素を考慮した自動要約手法を提案する．提案手法では，すべてのスレッドに対して階層的クラスタリングを行い，各クラスタを対象にグラフベースの手法である LexRank を用いて重要文を抽出する．さらに，アンケートによる評価実験を通して，提案手法が議論掲示板において有効であることを示す．

1. はじめに

Web の急速な発展により，膨大な量かつ多岐にわたる情報が世の中にあふれている．そのような状況の中で，我々は享受できるあらゆる情報に目を通し，内容を吟味した上で自分が取得すべき情報を取捨選択することはかなりの時間がかかることから，自然言語処理技術を応用した自動要約へのニーズは年々高まってきている．例えば，MMR(Maximal Marginal Relevance)[1] という冗長性排除の指標と併用し，潜在的な情報によって類似文が現れない要約を生成する手法が提案されている．一方，これらの文書要約は，新聞記事や学術論文など，単一もしくは数人の著者が執筆した文書を対象にした既存研究がほとんどである．

現存する多様なメディア形態の中でも，電子掲示板はブログなど個人で運営している小規模なものや，2 ちゃんねるや SNS に代表される不特定多数のユーザが参加する大規模なものがでてきている．使途や目的はそれぞれ異なるが，とりわけ Web 上で議論を行うことを目的とした掲示板を議論掲示板と呼ぶ．議論掲示板では，議論が深まるにつれて話題が多様化する傾向が見られる．例えば，あるユーザは新たな話題を持ち出し，あるユーザは他のユーザに返信を行うなどして，時間が経過するとともに掲示板における多種多様な意見が織りなす全体構造がより複雑なものへと変化する．そのような場面を想定したとき，ユーザはすべての変化にアンテナを張ることは難しいため，これまでに行われた議論の要約をユーザに提示することが有効である．

本研究では，議論掲示板のスレッドを対象に要約を自動的に生成し，最終的には参加ユーザに提示することによって，議論の内容理解を支援する．議論掲示板の一つである COLLAGREE[2] を題材として，議論掲示板の特徴を考慮した自動要約手法の提案および評価を行う．まず，複数のスレッドを対象として，時系列や返信関係，ユーザ情報といった議論掲示板が持つ特徴を考慮して階層的クラスタリングを行い，不要発言の除去を行う．次に，述語項構造と呼ばれる意味関係にある文節を抜き出すことによって文短縮を行う．その後，Web ページのランキング方式である PageRank[3] を複数文書要約に拡張した LexRank[4] により各文の持つ意見情報に基づいて，各クラスタから重要文を抽出する．最後に，要約文の並び替えを行い，要約文を出力する．また，要約結果に対して，評価アンケートを実施し，提案手法の有効性を示す．

以下に，本論文の構成を示す．第 2 章では自動要約手法に関する関連研究を示す．第 3 章ではスレッド構造と時系列を考慮した自動要約手法を提案する．その後，第 4 章で，評価実験結果と議論を行い，第 5 章で本論文のまとめを示す．

2. 関連研究

自動要約に関する研究は，新聞記事や学術論文など，単一もしくは数人の著者が執筆した文書を対象にした既存研

[1] 東京農工大学 工学部 情報工学科
Institute of Information and Computer Sciences, Tokyo University of Agriculture and Technology, Koganei, Tokyo,184-8588, Japan
[a)] kitagawa@katfuji.lab.tuat.ac.jp

究が多く，電子掲示板や SNS に代表される Web ページを対象とした既存研究は少ない．電子掲示板の関連研究として，前の発言者の発言を受けて，それに新しい情報を加えていくというスレッド構造に着目した松尾ら [5] の研究がある．ある投稿に出現する語が，返信先の投稿に出現しているならばその後を旧情報，そうでなければ新情報を担う語であると考え，電子掲示板を分析した結果に基づき，投稿がこれらの語をどれぐらい含んでいるかに基づいた要約手法を提案している．

また，議論掲示板のテキストデータは投稿された時刻情報を伴い, 時間経過に従って内容も新たな方向へと推移していくため，時系列データとみなすことができる．菊池ら [6] の研究では，この時系列テキストを対象に，クラスタ間類似度の計算時にクラスタ内文書の平均生起時刻の近さを考慮した階層的クラスタリングと複合語判定のための C-value 法を組み合わせ，話題の推移を表現したキーワード群を抽出している．羽鳥ら [7] は文の情報量を主たる基準としたこれまでの要約手法と異なり，PageRank アルゴリズムをを利用したグラフベースのモデルを構築した．提案手法において，手がかり語や返信関係，語彙的連鎖を新たな 3 要素として付加したアプローチをとり，各発言の重要文とスレッドのトピックを抽出している．北島ら [8] の手法では，PageRank を複数文書要約に発展させた LexRank を用いて，LDA にて推定したトピックを文間の類似度計算に代用している．また，MMR(Maximal Marginal Relevance)[1] という冗長性排除の指標と併用し，潜在的な情報によって類似文が現れない要約を生成することができる．しかし，これらの既存研究は議論掲示板の持つユーザや投稿時間という情報を活用した手法ではない．本論文では，議論掲示板の持つユーザや投稿時間という情報を活用し，自動要約を行う．

3. スレッド構造と時系列を考慮した自動要約手法

COLLAGREE[2] をはじめとした議論掲示板では，一つのテーマに対し，それに関連する単一のテーマを扱った複数のスレッドから構成される．ここでいうスレッドとは，ある特定の話題・論点に関する 1 つのまとまりを指す．例えば，図 1 のように，スレッドを立てたユーザの発言が親意見となり，他のユーザが子意見として親意見に返信し，その子意見に対して孫意見が存在する場合もある．

これらの議論掲示板のスレッドの集合に対して，本論文では，自動的に要約文を作成する手法を提案する．提案手法における処理の流れは，以下の通りである．

Step 1. 階層的クラスタリング
Step 2. 不要発言除去
Step 3. 文短縮
Step 4. 重要文抽出

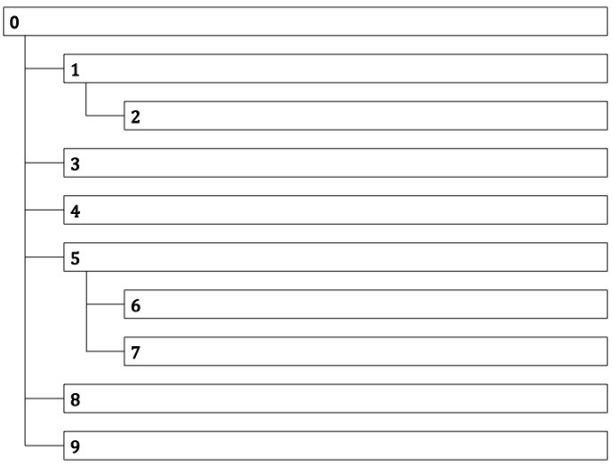

図 1 COLLAGREE のスレッドの概観

Step 5. 要約文の並び替え

3.1 階層的クラスタリング

スレッドを対象に凝集型の階層的クラスタリングを用いて，スレッド内部のさらに細分化した話題ごとに意見を分類する．クラスタリングを行うことにより，議論掲示板で行われる同様の内容で行われている発言群を発見し，決められた文字数での要約における網羅率を向上させる．文間の類似度計算には，以下に示すコサイン類似度を用いる．

$$sim(u,v) = \frac{\vec{u}\cdot\vec{v}}{|\vec{u}|\cdot|\vec{v}|} \tag{1}$$

$\vec{u},\vec{v}$ はそれぞれ文 u,v のベクトルを表す．文ベクトルは，対象となる文に対して，MeCab[9] により形態素解析を行い，文書から抽出した名詞と動詞に基づき，Paragraph Vector[10] を用いて生成したベクトルとする．

クラスタ間の類似度計算には群平均法を採用し，それぞれのクラスタに属するデータのすべての組み合わせの類似度の平均を，2 つのクラスター間の類似度とする．クラスタ数は，Upper Tail 法 [11] と呼ばれる停止規則に基づいて決定する．Upper Tail 法は，大きさ n の標本に対して，2 つのクラスタを併合する際の基準となるクラスタ間類似度の値が $n-1$ 個あることを利用する．最終的に 1 つのクラスタが形成されるときの基準値 α_1 から始めて，降順に α_{n-1} までの基準値がある．このとき，$j=1$ から始めて以下の条件を満たすまで j を増加させる．

$$\alpha_j \leq \bar{\alpha} + ks_\alpha \tag{2}$$

$\bar{\alpha}$ と s_α はそれぞれ α の分布の平均と不偏分散の平方根をとったものであり，j はクラスタ数を表している．k は定数であり，本手法では $k=1$ とする．

スレッドの最初の発言は，そのスレッドで話し合いたい議題の導入と，その議題について他のユーザに考えるきっかけを与えるための問題提起の 2 つの部分から構成されていることが多い．したがって，要約の初めにスレッド主の

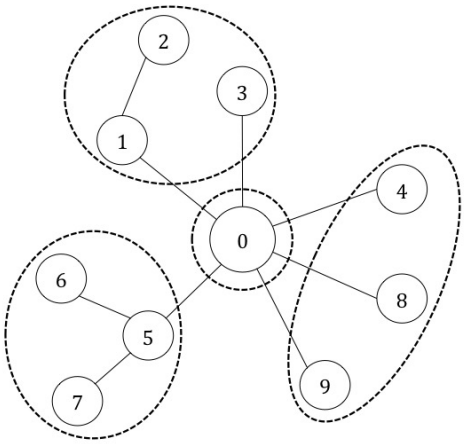

図 2 クラスタリングの例

意見を取り入れるべきである．階層的クラスタリングを行うことによって，他の意見が混在するクラスタに分類されないように，スレッドの最初の発言は単体で 1 つのクラスタを形成することとする．

図 1 のスレッドに含まれる意見を階層的クラスタリングした例を図 2 に示す．図 2 が示すように，ノードが各発言，エッジが発言間の関係を示しており，クラスタリングの結果 4 グループに分割された．

議論掲示板は，返信関係から成るスレッド構造を持ち，時系列により発言が着目する話題も移り変わっていく．そこで，参加しているユーザによって意見の偏りや傾向があると仮定し，階層的クラスタリングを行う場合に文間のコサイン類似度に加えて，投稿時刻の近さ・返信関係・同一ユーザ度の 3 要素を考慮する．以下に，投稿時刻の近さ・返信関係・同一ユーザ度の定義を示す．

投稿時刻の近さ

あるスレッドにおいて，互いに投稿時刻が近い意見は，同じ話題に言及している可能性が高いと考えられる．そこで，菊池らの手法 [6] を参考にして，投稿時刻の近さを考慮した減衰関数

$$w_{\mathrm{time}}(u,v) = \exp\bigl(-\alpha_{\mathrm{time}}(t_u - t_v)^2\bigr) \tag{3}$$

によって重み付けを行う．ただし，t_u, t_v は文 u, v の投稿時刻，α_{time} は定数である．このとき，式 (1) から文間の類似度は以下のようになる．

$$sim'(u,v) = sim(u,v) \cdot w_{\mathrm{time}}(u,v) \tag{4}$$

返信関係

スレッドは，ある特定の話題・論点に関する各発言の返信関係により成り立っており，スレッド内の返信関係にある発言には何らかの共通点がある．これらの返信関係をできるだけ保持したままクラスタリングを行うために，以下で定義する式によって返信関係を考慮した重み付けを行う．

$$w_{\mathrm{reply}}(u,v) = \begin{cases} \alpha_{\mathrm{reply}} & \text{返信関係にある} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

ただし，α_{reply} は定数を表す．このとき，式 (4) は以下のように更新される．

$$sim'(u,v) = \bigl(sim(u,v) + w_{\mathrm{reply}}(u,v)\bigr) \cdot w_{\mathrm{time}}(u,v) \tag{6}$$

同一ユーザ度

同じユーザによって発言された意見は，同じような考えを前提にして，投稿されている可能性があると考えられる．例えば，職業や立場，経験などの背景が挙げられるが，このことを踏まえると，発言しているユーザごとにまとまってクラスタに分類されることが望ましい．そこで，以下で定義する式を用いて同一ユーザ度に関する重み付けを行う．

$$w_{\mathrm{user}}(u,v) = \begin{cases} \alpha_{\mathrm{user}} & \text{ユーザが同じ} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

ただし，α_{user} は定数を表す．このとき，式 (6) は以下のように更新される．

$$\begin{aligned} sim'(u,v) = \bigl(&sim(u,v) + w_{\mathrm{reply}}(u,v) \\ &+ w_{\mathrm{user}}(u,v)\bigr) \cdot w_{\mathrm{time}}(u,v) \end{aligned} \tag{8}$$

3.2 不要発言除去

返信関係の上に成り立つスレッド構造を持つ議論掲示板では，意見の初めに「こんにちは，～です.」や，「～さん，その意見に同感です.」など，要約には含めるべきではないと思われる文が多々見られる．このような他のユーザへの挨拶や賛同・反対に分類される文などの不要発言を，重要文抽出を行う前にあらかじめクラスタから取り除き，要約に含めるべき重要文として抽出されることを防ぐ．

要約に必要/不要という観点に着目し，線形二値分類器であるサポートベクトルマシン (SVM) を導入して，不要発言の除去を行う．実際には，COLLAGREE の過去データから，正例 200 文・負例 200 文，合計 400 文を人手で選出し，訓練データとしている．さらに，要約には不要とみなされる文の特徴として，文字列長が短いことが挙げられる．分類精度を高めるために，Paragraph Vector によって生成した文書ベクトルに，文の文字列長情報を 1 次元加える．以上のように，文書ベクトルを用いて，訓練データから SVM を学習させ，クラスタに含まれるすべての文を対象に要約に必要な文と不要な文に分類する．

3.3 文短縮

述語項構造に着目した文短縮を行う．述語項構造とは，何らかの事態を表す述語に対して，それに不可欠な項の関係を指す．文字数に制約のある要約では，たとえ同じ分量でも包括している情報量が多い要約が生成されることが好ましい．そこで，あらかじめ文を短縮しておくことで一文

あたりの文字数が減らし，多くの文を重要文として抽出する．クラスタに含まれる文を対象に JUMAN/KNP[12] を用いて述語項構造解析を行う．そして，同定した述語項構造から，その関係にある文節のみを元の文から取り出し，文を再構築することで一文を短縮する．

3.4 重要文抽出

クラスタ内の各発言に対して LexRank を計算し，重要度を計算する．その後，重要度の高い発言から重要文を抽出する．

LexRank

LexRank は，Erkan らによって提案された手法であり，PageRnak を複数文書要約に拡張したものである [4]．例えば，図 3 に示すグラフにおいて，ノードが対象となる文書群に含まれる文，エッジが文間類似度を表している．グラフ表現における固有ベクトル中心性の概念に基づいて，隣接するノードの重要度も考慮しながら文の重要度を計算する．

グラフにおけるエッジの重み付け手法に関して，ある閾値を設定し，類似度が閾値に満たないエッジを枝刈りする重みなしグラフを用いる LexRank と，正規化した類似度をそのままエッジに利用する重み付きグラフを用いる Continuous LexRank の 2 種類が存在する．本手法では，クラスタ内の文書群から重要文を選定することを目的としている．1 つのクラスタに含まれる総文数は比較的少なくなることも予想され，そのような条件下でも，文の重要度において互いに有意な差が生じることが望ましい．したがって，正規化した類似度をそのままエッジに利用する重み付きグラフを用いる Continuous LexRank を採用することにした．

クラスタに含まれる文の重要度を計算するため，文 u の重要度は以下の式で求められる．

$$p(u) = \frac{(1-d)}{N} + d \sum_{v \in adj[u]} \frac{sim''(u,v)}{\sum_{z \in adj[v]} sim''(z,v)} p(v) \quad (9)$$

N は対象としている文書における総文数，$adj[u]$ は文 u に隣接するノードの集合，d はある一定の割合で非隣接ノードにジャンプするための制動係数を表す．d の値は Erkan らに倣い，$d = 0.85$ とした．

意見間の関係に基づいた類似度計算

要約に含めるべき文は，内容が冗長にならない範囲で，文間に結束性が見られるほうが意味が通じやすいと考える．ここで，3.1 節の階層的クラスタリングによって得られたクラスタには複数の意見が含まれており，2 文間の類似度をそれぞれの文を包含する意見間の関係を考慮した値で表現する．羽鳥らの手法 [7] を参考に，以下の式で式 (9) 中の文間類似度の値に重みを付与する．

$$sim''(u,v) = sim(u,v) \cdot rel(u,v) \quad (10)$$

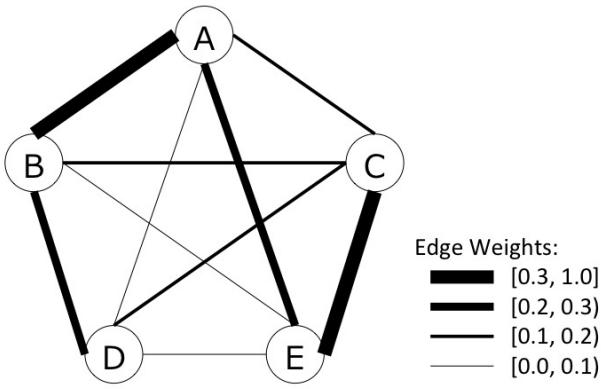

図 3 LexRank におけるグラフ表現

ただし，$rel(u,v)$ は文 u, v が同一意見に含まれるとき 2.0，返信関係にある意見に含まれるとき 1.5，全く関連のない意見に含まれるとき 1.0 とする．

3.5 要約文の並び替え

最後にユーザへの要約提示に際して，3.1 の階層的クラスタリングおよび 3.4 の重要文抽出により取り出した文を，前後関係に矛盾なく並び変える必要がある．まず，スレッドの最初の発言クラスタを，残りの発言から階層的クラスタリングを行うことによって形成したクラスタに統合し，各クラスタ内で指定した要約率を満たすように重要文を選ぶ．このような操作により，総文数が多いクラスタほど，抽出される重要文の数も多くなる．

最後に，議論掲示板にはそれぞれの発言に投稿時刻が情報として付与されているため，すべてのクラスタから抽出された重要文を投稿時刻の若い順に並び替える．投稿時刻による並び替えにより，要約全体の文脈が前後することを防ぐ．

4. 評価実験

4.1 実験設定

提案手法により生成した要約の有効性を評価するために被験者実験を行った．使用データには，2013 年 11 月に実施された名古屋市次期総合計画に関する 4 題（人権・環境・災害・魅力），および 2015 年 1 月に実施された愛知県がこれから目指す街づくりに関する 1 題の合計 5 つの COLLAGREE[2] の議論データの中から，投稿数が 10 以上のスレッドを 1 つずつ，合計 5 つのスレッドを用いた．

3.1 の階層的クラスタリングで述べた式 (3)，(5)，(7) 中のパラメータについては，$\alpha_{\text{time}} = 0.1$，$\alpha_{\text{reply}} = 0.1$，$\alpha_{\text{user}} = 0.1$ とする．また，要約率を変化させたとき，それが要約の質にどのように影響を及ぼすのかを調査するために，要約率 50%・25%・10%の 3 パターンに設定した．

本実験は 9 人の学生を対象に実験を行った．浅原ら [13] の人手による主観評価指標から本手法の評価に適するように変更している．以下が本実験で使用した，各要約文に対する評価項目である．

1. **可読性**　要約に意味が通じない文が含まれていないか
2. **非冗長性**　同じ情報が繰り返されていないか
3. **内容理解**　要約を読んで原文書の大意を把握できるか
4. **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
5. **順序の正確さ**　要約に含まれる文が前後していないか

各評価項目に対して，要約前のスレッドを提示した上で，A(とても良い)-E(とても悪い) の 5 段階評価を行う．

4.2 実験結果

図 4-6 はアンケートによる被験者実験の結果を示しており，それぞれ要約率を 50%・25%・10%とした場合の評価の平均割合を示している．全体的な評価に着目すると，要約率が 50%の場合が良い評価を示す A・B が最多となり，逆に要約率が低くなるにつれて悪い評価が増える傾向にある．より詳細に比較すると，「可読性」「非冗長性」「順序の正確さ」には，要約率の変化に対して，それほど大きな差異は見られなかったが，「内容理解」「網羅性」に関しては，要約率が低くなるにつれて悪い評価を表す D・E が増えている．

このような結果の理由として，要約率が高いほど要約に冗長な文が含まれる確率が高くなるため，同じ情報を含む要約を生成していないことが考えられる．一方，「内容理解」「網羅性」における A, B の割合は，要約率が 25%の場合は内容理解, 網羅性とも約 30%，要約率が 10%の場合はそれぞれ約 20%と約 5%であった．しかし，C と回答した被験者を考慮すると，要約率が 25%の場合は全体の 8 割程度であることから，概ね意味が通じ，かつ重要な情報を網羅した要約を生成したと考えられる．したがって，議論掲示板にて本手法を導入する場合を想定したとき，要約前のスレッドの内容を損ねない要約文をユーザに提示するには，要約率の下限を 25%程度に設定するのが妥当であると考えられる．

5. おわりに

本論文では，COLLAGREE という議論掲示板のスレッドを対象とし，要約文の自動生成手法を提案した．提案手法では，まず，文の内容に加えて投稿時刻の近さ，返信関係，同一ユーザ度の 3 要素を考慮した階層的クラスタリングを行う．次に，各クラスタから不要発言除去と文短縮を行った上で，Continuous LexRank に文の持つ意見情報を付加した重要文抽出を実行することで，議論掲示板に特化した自動要約を行う．また，被験者にシステムの要約結果を提示し，評価アンケートを実施することで，提案手法の有用性を確認した．特に，非冗長性の観点では，優れていることが確認された．一方，低い要約率での内容理解と網羅性の観点において，要約前のスレッドの内容を損ねている問題点が明らかになった．

今後の課題として，文短縮のさらなる改良が考えられる．文短縮では，述語項構造とは関係のない文節を取り除くことに終始したが，圧縮率の高い文短縮手法を考案することでスレッドの網羅性やユーザのさらなる内容理解につながると考えられる．また，提案手法は様々なフェーズから成り立っているため，各フェーズの効果に関してより詳細な比較実験および考察する必要がある．

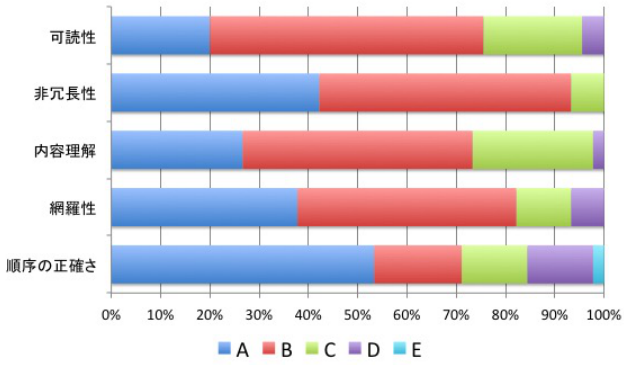

図 4　要約文に対する各評価項目の平均値 (要約率 50%)

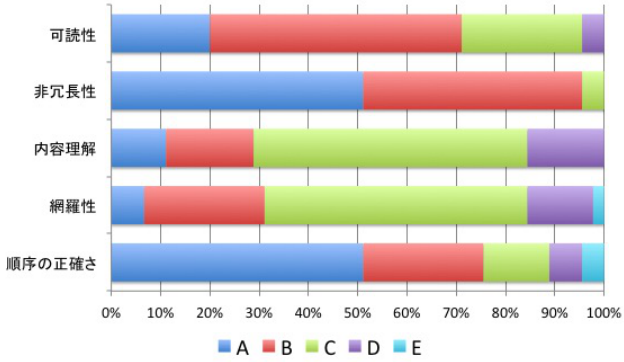

図 5　要約文に対する各評価項目の平均値 (要約率 25%)

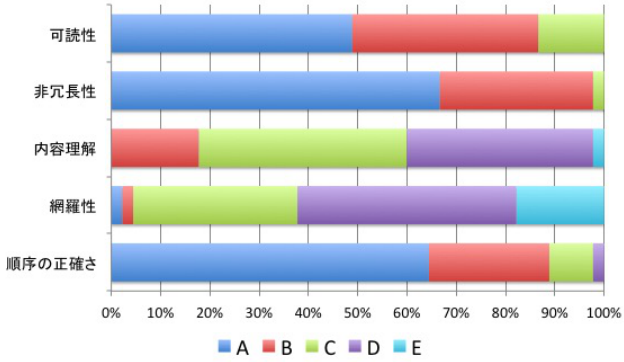

図 6　要約文に対する各評価項目の平均値 (要約率 10%)

謝辞

本研究は，JST，CREST の支援を受けたものである．

参考文献

[1] Goldstein, J., Mittal, V., Carbonell, J. and Kantrowitz, M.: Multi-document summarization by sentence extraction, *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4*, Association for Computational Linguistics, pp. 40–48 (2000).
[2] 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:～ 議論プロセスの弱い構造化による意見集約支援 ～，日本経営工学会論文誌，Vol. 66, No. 2, pp. 83–108 (2015).
[3] 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).
[4] 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).
[5] 松尾　豊，大澤幸生，石塚　満：電子掲示板における会話からのトピックの発見と要約，人工知能学会全国大会論文集，Vol. 16, pp. 1–4 (2002).
[6] 菊池匡晃，岡本昌之，山崎智弘：階層型クラスタリングを用いた時系列テキスト集合からの話題推移抽出，日本データベース学会論文誌，Vol. 7, No. 1, pp. 85–90 (2008).
[7] 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 (NLP2010)，pp. 290–293 (2010).
[8] 北島理沙，小林一郎：潜在トピックを考慮したグラフ表現に基づく複数文書要約，知能と情報，Vol. 25, No. 6, pp. 914–923 (2013).
[9] MeCab: Yet Another Part-of-Speech and Morphological Analyzer (2006). http://mecab.sourceforge.netz/.
[10] Le, Q. V. and Mikolov, T.: Distributed representations of sentences and documents, *The 31st International Conference on Machine Learning*, pp. 1188–1196 (2014).
[11] Mojena, R.: Hierarchical grouping methods and stopping rules: An evaluation, *The Computer Journal*, Vol. 20, No. 4, pp. 359–363 (1977).
[12] 河原大輔，黒橋禎夫：自動構築した大規模格フレームに基づく構文・格解析の統合的確率モデル，自然言語処理，Vol. 14, No. 4, pp. 67–81 (2007).
[13] 浅原正幸，杉　真緒，柳野祥子：BCCWJ-SUMM：『現代日本語書き言葉均衡コーパス』を元文書とした要約文書コーパス，第 7 回コーパス日本語学ワークショップ予稿集，pp. 285–292 (2015).