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

渋木 英潔†　森 辰則†

†横浜国立大学 大学院 環境情報研究院
E-mail: {shib,mori}@forest.eis.ynu.ac.jp

1　はじめに

文書中の重要な文を抽出して提示する抜粋型の要約には，MMR に基づく手法 [1] や，整数計画問題として解く手法 [2] など多くの手法が提案されている．その中に，TextRank[3] や LexRank[4] に代表されるグラフベースアルゴリズムに基づく手法が存在する．文書中の各文をノードとして，類似性や包含関係などの文間関係をリンクの重みとして表現する要約手法では，以下の 2 点が議論の対象となることが多い．一点目は，高村ら [5] のように，ある尺度に基づいてノードの重要度が計算された場合に，どのように冗長性を排除しつつ重要度が高いノードから優先的に被覆するかということに関する議論であり，二点目は，Kaneko et al.[6] のように，適切な重要度を計算するための尺度やグラフ構造に関する議論である．本稿では，後者に主眼を置いて議論を進める．

後者の議論に主眼を置いたグラフベースの手法に，Co-HITS-Ranking を用いた Hu et al.[7] の手法がある．Hu et al. の Co-HITS-Ranking では，以下の 2 つの仮説に基づいて，文層と文書層の 2 層の構造をもつグラフにおける重要度の計算を行っている．第一の仮説は,「クエリや重要な文（文書）と重くリンクされた文（文書）が重要な文（文書）である」という，文同士または文書同士といった同じ種類のノード間の関係に関する仮説であり，第二の仮説は,「重要な文書（文）に包含される（する）文（文書）が重要な文（文書）である」という，文書と文といった異なる種類のノード間の関係に関する仮説である．Hu et al. の手法では，文書層のノードと文層のノードという異層間の情報を統合することで精度の改善を行ったが，我々は以下の 3 つの疑問点に関する調査を行うことで，さらに改善のための議論ができるのではないかと考えた．

一点目は，Hu et al. が文や文書の概念を構成する概念単位 [8] として単語を用いており，Bag of Words（以降，BoW）による文間類似度の計算を行っている点である．単語を概念単位とした場合，トピックレベルの一致度を測るには十分であるが，命題などの，より詳細なレベルの一致度を測るには粒度が粗いことが多い．Hovy et al.[9] は，最小の意味的な単位として Basic Element を提案しており，Basic Element を用いることでより詳細なレベルの一致度を測ることができると考えられる．しかしながら，Hovy et al. は，要約を評価する単位として Basic Element を用いているが，要約生成における概念単位として Basic Element を用いてはいない．そこで，文間類似度の尺度を計算する単位として Basic Element を用いた場合，すなわち，BoW ではなく Bag of Basic Elements（以降，BoBE）による計算を行った場合に，どのような影響があるかを調査する．

二点目は，Hu et al. のグラフ構造が，文書層と文層の 2 層で構成されている点である．我々は，文献 [6] において，文書層，パッセージ層，文層，単語層の 4 層で構成されるグラフ構造を提案しており，例えば，BoW による文間類似度は，二つの文ノードが共通の単語ノードを包含するグラフ構造で表現することができる．したがって，文書層，文層に，単語層を加えた 3 層のグラフ構造を用いた場合に，どのような影響があるかを調査する．

三点目は，Hu et al. では，文書層におけるノード間の関係を示す尺度と，文間におけるノード間の関係を示す尺度が，同じ観点からの指標である点である．異層間の情報を統合することで精度の改善を行うという目的において，それぞれの層で扱う情報の質が類似したものである場合，統合による効果が薄くなるのではないかという不安がある．例えば，文書層ではトピックレベルの一致度を測るために BoW による類似度を，文層では命題レベルの一致度を測るために BoBE による類似度を，単語層では語義レベルの一致度を測るためにシソーラス距離による類似度をそれぞれ用い，そ

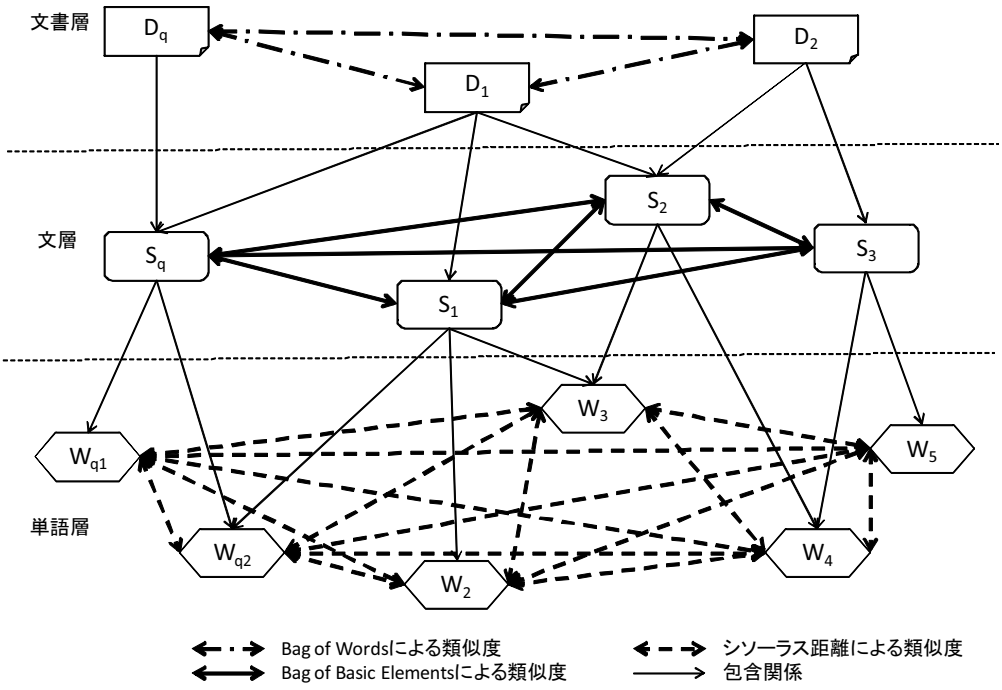

図 1: 提案手法のグラフ構造

れらの情報を統合した方が，多様な観点からの総合的な判断効果が得られるのではないかと考えた．そこで，各層におけるノード間の関係を示す尺度を独立して変更した場合に，どのような影響があるのかを調査する．

以上から，本稿では，文書層，文層，単語層の 3 層で構成されるグラフ構造において，各層におけるノード間の類似度に，BoW，BoBE，シソーラス距離を用いたグラフベースの重要文抽出手法を提案する．なお，Hu et al. はクエリ指向の要約を対象としている．本稿で提案するアルゴリズムはクエリ指向の要約に限定されるものではないが，Hu et al. の手法との比較を行う上で，4 節の実験ではクエリ指向の要約を対象としている．

2 基本的な考え方

図 1 に，提案手法で用いるグラフ構造を示す．文書層，文層，単語層の 3 層で構成されている．文書層のノード間類似度（リンクの重み）にどのような指標を用いるかであるが，文書という大きな言語単位において，依存関係などの構造を正確に捉えるのは困難であると考えたため，トピックレベルの一致度に相当すると思われる BoW による類似度を用いることとした．一方，文層のノード間類似度では，依存関係による構造を厳密に捉えた方が良いと考えたため，命題レベルの一致度に相当すると思われる BoBE による類似度を用いることとした．単語層のノード間類似度では，そもそも 1 語であるため BoW や BoBE による類似度は意味をなさない．そのため，シソーラス距離による類似度を用いることで，語義レベルの一致度を計算することとした．

3 提案手法

ノードの重要度を計算するアルゴリズムは，基本的に Hu et al.[7] の Co-HITS-Ranking と同じものである．2 つのノード N_1 と N_2 の間の，BoW によるノード間類似度 $Sim_{BoW}(N_1, N_2)$，BoBE によるノード間類似度 $Sim_{BoBE}(N_1, N_2)$，シソーラス距離によるノード間類似度 $Sim_{TD}(N_1, N_2)$ は，以下の式 (1-3) でそれぞれ計算される．

$$Sim_{BoW}(N_1, N_2) = \frac{cbow(N_1, N_2)}{|bow(N_1) \cup bow(N_2)|} \quad (1)$$

$$Sim_{BoBE}(N_1, N_2) = \frac{cbobe(N_1, N_2)}{|bobe(N_1) \cup bobe(N_2)|} \quad (2)$$

$$Sim_{TD}(N_1, N_2) = \frac{D_{TD} - depth(N_C)}{D_{TD}} \quad (3)$$

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，$bow(N)$ はノード N のテキストに含まれる単語の集合であり，$cbobe(N_1, N_2)$ 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，$bobe(N)$

表 1: 第一の実験結果：BoW による類似度と BoBE による類似度の比較

着目言明	正解数	BoW		BoBE	
レーシック手術は安全である	755	64	(8.5%)	240	(31.8%)
レーシック手術は痛みがある	296	2	(0.7%)	36	(12.2%)
無洗米は水を汚さない	677	5	(0.7%)	114	(16.8%)
無洗米はおいしい	783	7	(0.9%)	71	(9.1%)
アスベストは危険性がない	188	6	(3.2%)	37	(19.7%)
キシリトールは虫歯にならない	1,183	36	(3.0%)	167	(14.1%)

はノード N のテキストに含まれる Basic Element の集合である．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 のシソーラス上の共通上位ノード，$depth(N)$ はシソーラス上のルートノードからノード N までの距離である．

4 実験

1 節で述べた 3 つの疑問点をそれぞれ調査するための三つの実験を行う．第一の実験は，文間類似度の尺度として BoW と BoBE を用いた場合に，どのような影響があるかを調査するものである．文書層と単語層からの情報を統合せずに，文層内のリンクのみを用いて重要度の計算を行う．第二の実験は，文書層や単語層の情報を統合した場合に，どのような影響があるかを調査するものである．文層のノード間類似度を BoBE を用いた場合に固定し，BoW によるノード間類似度を用いた文書層と統合した場合，BoBE によるノード間類似度を用いた文書層と統合した場合，シソーラス距離によるノード間類似度を用いた単語層と統合した場合の比較を行う．第三の実験は，三層で構成されるグラフ構造において，文書層と文層のノード間類似度に，BoW と BoBE の同種または異種の組み合わせを用いた場合に，どのような影響があるかを調査するものである．

実験データとして，Nakano et al.[10] で構築されたサーベイレポートコーパスを用いた．サーベイレポートコーパスは，情報信憑性判断支援のための要約を目的とした手法の評価・分析用コーパスである．情報信憑性判断支援のための要約は，利用者が信憑性を判断したい着目言明に関する Web 文書集合中の記述を抽出・整理して提示する，クエリ指向の抜粋型複数文書要約であり，サーベイレポートコーパスには，着目言明をクエリとして検索された Web 文書集合が収録されており，作業者により着目言明に関連する文と判断された文が重要文として Web 文書集合から網羅的に抽出されている．重要文の抽出は 4 名の作業者により行われており，本実験では 4 名中 1 名以上が重要文と判断した文を抽出すべき正解の文とした．サーベイレポートコーパスに収録されている着目言明と正解文の数を表 1 の第 1 列と第 2 列にそれぞれ示す．

評価尺度には R 精度を用いた．R 精度とは，正解データ数が R 個の場合，結果の上位 R 個にある正解データの割合である．R 精度を見ることで実験データ中における正解データをどれだけ上位にすることができているかを評価することができる．

各実験の結果を表 1 から表 3 にそれぞれ示す．表 1 の結果から，文間類似度に BoBE を用いることで大きく精度が向上したことが示された．表 2 の結果から，文書層のノード間類似度に何を用いるかは関係なく，文書層のみを統合しても精度の向上は見られなかったことが示された．一方，単語層を統合した場合は，一部の着目言明を除いて，大きく精度が向上したことが示された．文書層の統合が意味をなさなかったことは，BoBE を用いた文間類似度が非常に有効に働いており，それより粗い文書単位やトピックレベルの情報が意味をもたなかったことが原因と考えられる．一方，単語層の語義レベルの一致度は，BoW や BoBE と全く異質の情報であったため，統合することで精度の向上に寄与できたと考えられる．表 3 の結果から，文書層が単語層と共に統合された場合には，一部の着目言明において精度の向上があったことが分かる．しかしながら，その差が小さいこと，精度が低下した着目言明も存在することから，誤差の範囲と思われる．詳細な分析は今後の課題である．

5 まとめ

本稿では，文書層，文層，単語層の 3 層で構成されるグラフ構造において，各層におけるノード間の類似

表 2: 第二の実験結果：文書層の統合と単語層の統合の比較

着目言明	文書層:統合 (BoW) 単語層:非統合		文書層:統合 (BoBE) 単語層:非統合		文書層:非統合 単語層:統合	
レーシック手術は安全である	238	(31.5%)	238	(31.5%)	410	(54.3%)
レーシック手術は痛みがある	36	(12.2%)	36	(12.2%)	58	(19.6%)
無洗米は水を汚さない	114	(16.8%)	114	(16.8%)	171	(25.3%)
無洗米はおいしい	71	(9.1%)	71	(9.1%)	164	(20.9%)
アスベストは危険性がない	37	(19.7%)	37	(19.7%)	67	(35.6%)
キシリトールは虫歯にならない	95	(8.0%)	95	(8.0%)	88	(7.4%)

表 3: 第三の実験結果：文書層と文層における BoW と BoBE の同種または異種の組み合わせの比較

着目言明	文書層:BoW 文層:BoW		文書層:BoW 文層:BoBE		文書層:BoBE 文層:BoW		文書層:BoBE 文層:BoBE	
レーシック手術は安全である	63	(8.3%)	**410**	(54.3%)	63	(8.3%)	**410**	(54.3%)
レーシック手術は痛みがある	2	(0.7%)	**67**	(22.6%)	2	(0.7%)	65	(22.0%)
無洗米は水を汚さない	6	(0.9%)	156	(23.0%)	5	(0.7%)	**165**	(24.4%)
無洗米はおいしい	7	(0.9%)	**167**	(21.3%)	7	(0.9%)	162	(20.7%)
アスベストは危険性がない	4	(2.1%)	**63**	(33.5%)	4	(2.1%)	56	(29.8%)
キシリトールは虫歯にならない	36	(3.0%)	**109**	(9.2%)	36	(3.0%)	**109**	(9.2%)

度に，BoW，BoBE，シソーラス距離を用いたグラフベースの重要文抽出手法を提案した.

参考文献

[1] J. Goldstein, V. Mittal, J. Carbonell, M. Kantrowitz, “Multi-document Summarization by Sentence Extraction,” In Proc. of the 2000 NAACL-ANLPWorkshop on Automatic Summarization, vol.4, pp.40-48, 2000.

[2] 高村大也, 奥村学, “施設配置問題による文書要約のモデル化,” 人工知能学会論文誌, vol.25, no.1, pp.174–182, 2010.

[3] R. Mihalcea, “Graph-based ranking algorithms for sentence extraction, applied to text summarization,” In Proc. of the ACL 2004 on Interactive Poster and Demonstration Sessions (ACLdemo ’04), 2004.

[4] G. Erkan and D. R. Radev, “LexRank: Graph-based Lexical Centrality As Salience in Text Summarization,” Journal of Artificial Intelligence Research, Vol.22, pp.457–479, 2004.

[5] 高村大也, 奥村学, “最大被覆問題とその変種による文書要約モデル”, 人工知能学会論文誌, Vol.23, No.6, pp. 505–513, 2008.

[6] K. Kaneko, H. Shibuki, M. Nakano, R. Miyazaki, M. Ishioroshi, T. Mori, “Mediatory Summary Genaration: Summary-Passage Extraction for Information Credibility on the Web,” In Proc. of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 23), 2009.

[7] P. Hu, D. Ji, C. Teng, “Co-HITS-Ranking Based Query-Focused Multi-Document Summarization,” In Proc. of the 6th Asia Information Retrieval Societies Conference (AIRS 2010), pp 121–130, 2010.

[8] E. Filatova and V. Hatzivassiloglou, “A Formal Model for Information Selection in Multi-sentence Text Extraction,” In Proc. of the 20th International Conference on Computational Linguistics (COLING ’04), 2004.

[9] E. Hovy, C. Lin, L. Zhou, J. Fukumoto, “Automated Summarization Evaluation with Basic Elements,” In Proc. of the Fifth Conference on Language Resources and Evaluation (LREC 2006), 2006.

[10] M. Nakano, H. Shibuki, R. Miyazaki, M. Ishioroshi, K. Kaneko. T, Mori, “Construction of Text Summarization Corpus for the Credibility of Information on the Web”, Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010), 2010