多層ネットワーク型 TextRank による根拠関係を考慮した重要パッセージ抽出

永井 隆広[†]　金子 浩一[‡]　渋木 英潔[§]　中野 正寛[‡]
宮崎 林太郎[‡]　石下 円香[§]　森 辰則[§]
[†]横浜国立大学 工学部　[‡]横浜国立大学 大学院 環境情報学府　[§]横浜国立大学 大学院 環境情報研究院
E-mail: {nagadon,kaneko,shib,nakano,rintaro,ishioroshi,mori}@forest.eis.ynu.ac.jp

1　はじめに

Web 上に存在する情報は，ブロードバンド化の進展やブログ等の普及に伴い，爆発的に増加し続けている．これらの情報の中には出所が不確かな情報や利用者に不利益をもたらす情報などが含まれており，信頼できる情報を利用者が容易に得るための技術に対する要望が高まっている．しかしながら，情報の内容の真偽や正確性を検証することは困難である上に，その情報が意見などの主観を述べるものである場合には，利用者により考え方や受け止め方が異なることから，その真偽や正確性を検証することはさらに困難なものとなる．そのため，情報の信憑性は，最終的に個々の情報利用者が判断しなければならず，利用者による信憑性の判断を支援する技術の実現が優先して解決すべき課題であると考える．我々は，主観的な意見や評価だけでなく，疑問の表明や客観的事実の記述を含めたテキスト情報を広く言明と呼ぶこととし，ある言明集合における個々の言明の相対的な位置づけを提示することで利用者の情報信憑性の判断を支援することを目指している．

村上ら [4][7] は利用者の情報信憑性判断を支援するために，言明間に存在する類似，対立，含意等の論理的関係を解析してマップ化する言論マップの生成を行っている．言論マップによって，利用者は言明間の構造を把握することが容易になるので言論マップによる情報信憑性判断支援は有効であると考えられる． Web 上に現れる実際の文 (以下，実文) を 2 つ取り出した時に，これらが全体として同義や矛盾関係を満たすことは実際には稀であるので，言論マップでは実文から適当な構成要素を言明として取り出し得られた縮約された言明同士の間に意味的関係を付与している．

人間が文章の信憑性を判断する場合，対立や根拠といった骨子となる文間の関係だけでなく，ニュアンスの伝わり方の違いなど，個人の感性に影響される微妙な表現も考慮されることが多い．しかし，実文から言明を取りだすことで，このような微妙な表現や言明の成立する前提条件といった情報信憑性判断に役立つ情報が切り落とされてしまう恐れがある．そういった微妙な表現を現在の技術で正確に処理することは困難なので，可能な限り原文書の表現を保持したまま提示することが最適と考えられる．そこで我々は，言論マップ生成タスクの出力を前提として，重要パッセージ抽出を用いた抜粋型の要約による情報信憑性判断の支援を検討している．

本稿では，情報信憑性判断を支援するための要約システムの要素技術の一つである，重要パッセージ抽出について述べる． 2章では，情報信憑性判断のための要約と一般的な目的の要約との差異に関して検討を行い，我々がなぜ多層ネットワーク型の TextRank による重要パッセージ抽出を行うかを説明する． 3章では，利用者の情報信憑性判断を支援するための要約システム (サーベイレポート) についてその概要を述べる． 4章では，同要約システムの要素技術の一つである，根拠関係を考慮した重要パッセージ抽出の手法について説明する．また，提案手法について実験を行い，その結果を考察する． 5章はまとめである．

2　情報信憑性判断のための要約

2.1　一般的な目的の要約との差異

情報信憑性判断のための要約は， Web 上の複数文書を対象とした報知的要約の一種であるが，一般的な目的の要約とは以下の点で異なる．一般的な目的の要約では，入力として文書集合が与えられるが，情報信憑性のための要約では，「ディーゼル車は環境に悪いか？」といった，利用者が信憑性を判断したい言明が与えられる．これは， Web 検索エンジンにクエリを与えるように，我々のシステムに利用者が判断したい言明を与えることを想定している．

現在の自動要約の技術には，主に重要文抽出による要約と文圧縮による要約等がある [8]．これに対して，情報信憑性のための要約では，可能な限り原文書の表現を保持したまま提示することが最適と考えられることから，重要パッセージ抽出による抜粋型の要約が中心技術となると我々は考えている．我々は言論マップ生成システムと連携し，重要パッセージ抽出による要約を用いて図 1のようなサーベイレポートを生成することで，利用者の情報信憑性判断を支援することを目指す．サーベイレポートについては 3章で述べる．

2.2　重要パッセージ抽出と多層ネットワーク構造

Mihalcea ら [3] は PageRank[1] を自然言語処理に適用した TextRank による重要文抽出を提案した． TextRank はグラフ構造に基づいたランキングアルゴリズムであり，頂点となるテキストの断片について，その局所的情報ではなくグラフ構造全体から得られるテキスト全体に関わる大域的な情報をもとに頂点の重要度を決定する． TextRank ではリンクの重みとして類似度を用いており，意味的関係は考慮されていない．

また， Radev ら [5] は複数文書中の文間の関係解析に Cross Document Theory (CST) を提案した． CST は RST[6] に基づく談話構造解析を文書横断構造解析

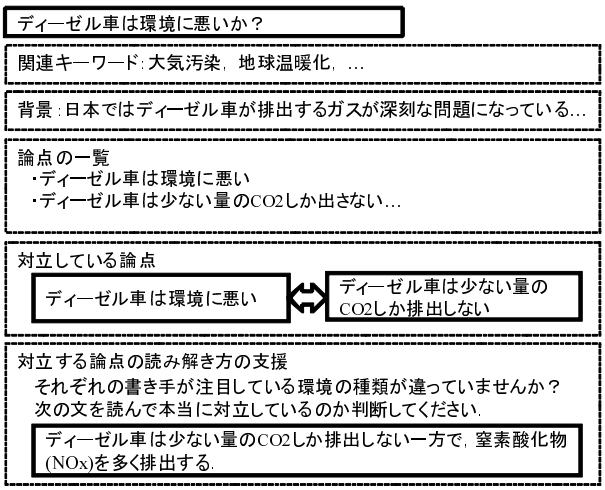

図 1: サーベイレポートの例

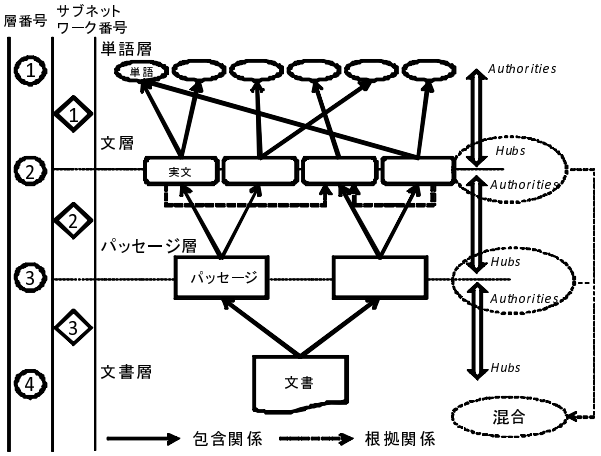

図 2: 多層ネットワーク構造の例

に拡張するものであり，18 種類の意味的関係がコーパス中で定義された．我々は CST 等のような論理関係を TextRank に反映させることで抽出精度を向上させることが出来ると考えている．

ここで図 2のような単語，文，パッセージ，文書の各層に分かれた多層ネットワーク構造と各層毎に別々のネットワークを用いた場合を比較・検討する．各層毎に別々のネットワークを用いた場合，重要な文を含むパッセージは重要であるというような各層間の関係を反映させることができない．一方，多層ネットワークを用いた場合には，各層間に包含関係のリンクを張ることでそのような関係を考慮することができる．また，我々が最終的な目標とするサーベイレポート (図 1) では，重要パッセージ抽出に加え，関連キーワードとして重要な単語を抽出を行う予定である．各層毎に別々のネットワークを用いた場合，単語層とパッセージ層のそれぞれについて Rank 付けを行う必要が生じる．一方，多層ネットワークを用いた場合には，一度の処理で全ての要素にスコアが付加されるので，各層毎にスコアを取り出すことにより，容易に各層毎のランキング付けが可能となる．このような利点より我々は多層ネットワーク構造を導入し，これに対する TextRank の計算手法を検討することとした．さらに，この多層ネットワーク上のリンク情報として，論理関係を導入することとした．多層ネットワークに TextRank を適用する手法については 4章で述べる．

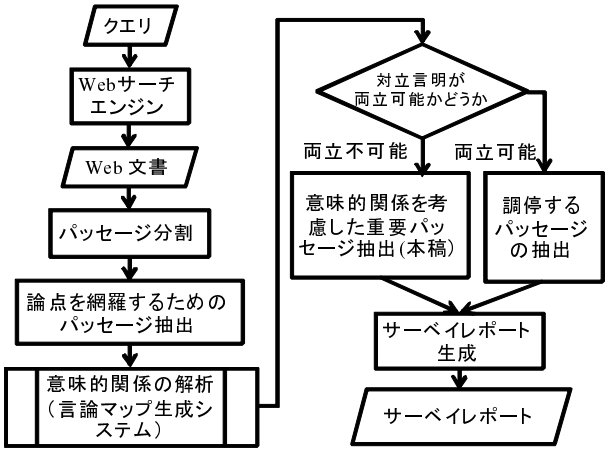

図 3: 提案システムの処理の流れ

3 サーベイレポート作成のための提案システムの概要

我々が提案する情報信憑性判断の支援のためのサーベイレポート (図 1) は，大きく 5 つの部分に分けられる．第一と第二の部分は，着目言明に関連するキーワードと背景となるイベントである．第三の部分は，着目言明における論点の一覧であり，ここまでの部分が着目言明のトピック全体を俯瞰するための要約となる．第四の部分は，対立関係に焦点を当てた注目すべき論点の一覧であり，第五の部分は，それらの対立点をどのように解釈すべきかに関する記述となっている．

図 3は，我々が提案するシステムで想定される処理の流れである．システムは与えられたクエリに対して，検索エンジン TSUBAKI[1]を使って Web 文書を検索する．パッセージ単位での抽出を行うため，検索された Web 文書を，HTML により記述された文書構造を考慮してパッセージに分割する．

提案システムでは言論マップ生成システムと連携し，言明間の関係の解析結果を取得する予定である．言論マップ生成システムに解析を依頼する前に，解析時間の短縮や後の処理の効率化のため，テキストの量を減らすための大まかな絞り込みを行う．この絞り込みでは，論点が網羅されたパッセージを残しつつ，不要なテキストを削減することを目指す．

言論マップ生成システムによって解析された対立関係について，システムはそれらの言明が両立可能となるような視点 (調停視点)[2] を探す．調停視点が見つかった場合，調停要約 [2] として利用者に提示する．見つからなかった場合は，各言明の根拠等を含むパッセージを探し，言明の真偽の判断材料として提示する．その場合には，本稿で述べている多層ネットワーク型 TextRank による根拠関係を考慮したパッセージ抽出を行う．

[1] http://tsubaki.ixnlp.nii.ac.jp

4 根拠関係を考慮した重要パッセージ抽出

4.1 HITS アルゴリズム

HITS アルゴリズムとは Web ページのランキング手法の一つであり，Hub と Authority の 2 種類のスコアを重要度に用いる．HITS アルゴリズムは，良い Hub からリンクを張られているページ (Authority) は良い Authority で，良い Authority にリンクを張るページ (Hub) は良い Hub であるという考えを基に，繰り返し計算していくことでランキング付けを行うアルゴリズムである．

Mihalcea ら [3] は，Web ページのランキング手法を抜粋型要約に応用する手法を TextRank として提案しており，HITS アルゴリズムによる TextRank も提案している．本来の HITS アルゴリズムでは，Hyperlink によるネットワーク構造を基本としているため，各ノード間の結合度に差異はなく，全てのリンクの重みは等しいものとしていた．しかしながら，TextRank では，類似度などのノード間の結合の強さに応じて，リンクの重みを考慮した計算を行っており，ノード V_i からノード V_j に重み w_{ij} のリンクが張られている場合，以下の式 (1),(2) に従って，それぞれ，Authority と Hub の値が計算される．

$$HITS_A^{t+1}(V_i) = \sum_{V_j \in In(V_i)} w_{ji} \cdot HITS_H^t(V_j) \quad (1)$$

$$HITS_H^{t+1}(V_i) = \sum_{V_j \in Out(V_i)} w_{ij} \cdot HITS_A^t(V_j) \quad (2)$$

$In(V_i)$ はノード V_i をリンク先とするリンクにおいてリンク元にあるノード集合を表し，$Out(V_i)$ はノード V_i をリンク元とするリンクにおいてリンク先にあるノード集合を表している．

4.2 多層ネットワークの導入

我々は，論理関係などの意味層にあるべき情報を実文層に写像するために図 2のような多層ネットワーク構造を用いることとした．図 2は，単語，文，パッセージ，文書といった粒度ごとの多層構造をしており，各層のノード間には，包含関係に基づいて有向リンクが張られている．また，意味的関係の一つとして根拠関係を図中の破線で示したような文ノード間のリンクとして表現することとした．

ここで，根拠となる言明は帰結に相当する言明の信憑性を利用者が判断する助けになると考えられる．そこで HITS アルゴリズムの考え方に習い，信頼できる帰結を含む根拠は信頼できる根拠であると考えれば，根拠を Hub，帰結を Authority として HITS アルゴリズムに適用できる．

さらに我々は，多くの重要文を含むパッセージは重要であり，多くの重要パッセージを含む文書は重要であるというように考える．つまり，ある粒度の文書断片の重要度は，その文書断片に含まれるより小さい文書断片の重要度と，その文書断片を含むより大きい文書断片の重要度の両方を考慮して計算する必要があると考える．その実現のために，HITS アルゴリズムによる TextRank を用いて，前者のより小さい文書断片の重要度を Authority，後者のより大きい文書断片の重要度を Hub とみなした計算を行い，両者を統合した値の高い順にランク付けして重要パッセージを出力することとした．

多層ネットワーク構造は 3 つのサブネットワークで構成されており，頂点の Hub と Authority のスコアはそれぞれのネットワークで計算される．層とサブネットワークの番号を図 2のように定義するとき，層 l における頂点 V_i の t 回目の繰り返し計算を行った時の Hub と Authority のスコアは以下の式で表わされる．

$$HITS_{A,l}^{t+1}(V_i) = \sum_{V_j \in In(V_i)} w_{ji} \cdot HITS_{H,l+1}^t(V_j) \quad (3)$$

$$HITS_{H,l}^{t+1}(V_i) = \sum_{V_j \in Out(V_i)} w_{ij} \cdot HITS_{A,l-1}^t(V_j) \quad (4)$$

本稿では，重み w の値を，包含関係のリンクは 1 に固定し，根拠関係のリンクは 0 以上としている．根拠関係のリンクの適切な重みについては 4.3節で検討する．

層 l における頂点の Authority は l 番目のサブネットワーク内で計算される．一方で，頂点の Hub は $l-1$ 番目のサブネットワークで計算される．異なるサブネットワークで計算された Hub と Authority は，次の式で混合され，一つの重要度として表わされる．

$$mHITS_{A,l}^t(V_i) = \alpha HITS_{A,l}^t(V_i) + (1-\alpha) HITS_{H,l}^t(V_i) \quad (5)$$

$$mHITS_{H,l}^t(V_i) = (1-\beta) HITS_{A,l}^t(V_i) + \beta HITS_{H,l}^t(V_i) \quad (6)$$

なお，α と β は係数で，今回の実験では共に 0.5 とした．この式によって混合された値を繰り返し計算の次の回で用いる．

4.3 評価実験

根拠関係を導入した多層ネットワーク構造の有効性を確認するために実験を行った．正解データには 4 人の作業者が作成した抜粋要約を用いた．抜粋要約に用いたトピックは「無洗米はおいしい」と「レーシック手術は痛い」の 2 トピックであり，この両方について 4 人の作業者がそれぞれ抜粋要約を作成した．実験では，パッセージは連続した 5 文とし，作業者が抜粋要約に使用した文を少なくとも 2 文含んでいるパッセージを正解とした．本節における実験では，抜粋要約に使用した文のみであるため，評価基準が厳しくなっている．「無洗米はおいしい」のトピックでは 534 パッセージ中，正解パッセージが 221 パッセージあり，「レーシック手術は痛い」のトピックでは 535 パッセージ中，正解パッセージは 178 パッセージ含まれていた．根拠関係のリンクは人手でつけられており，リンクの重みは 0 から 50 まで変化させた．我々は，R 精度と上位 30，50，100 パッセージ中の精度を調べた．R 精度とは，出力された重要度ランキングの上位から正解の数だけパッセージを抽出し，それらの中に含まれた正解数の割合によって順位付けの質を評価するものである．

図 4および図 5は実験の結果である．図中の正解パッセージの割合は，ランダムに抽出した場合の精度と見なすことができる．また，根拠関係のリンクの重みが 0

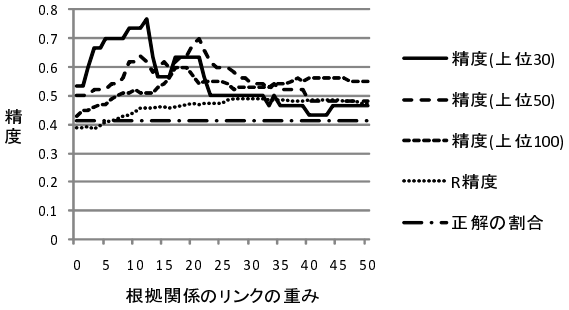

図 4: トピック「無洗米はおいしい」における精度

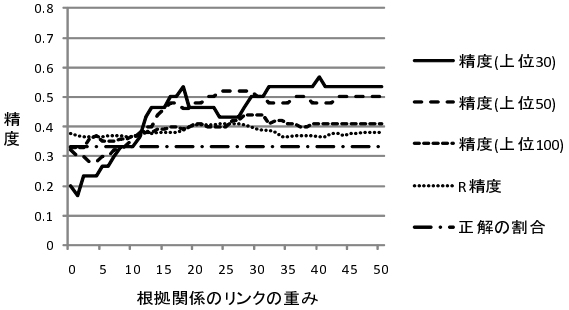

図 5: トピック「レーシック手術は痛い」における精度

のときが，従来の HITS ベースの TextRank による精度となる．より上位のパッセージの抽出精度が高いほど，我々の手法による順位づけの質が高いと判断できる．

上位 100，50，30 の精度については上位になればなるほど精度が向上しており，上位のパッセージの抽出精度が高くなっている．根拠関係のリンクの重みが最適な値になったとき，上位 30 の精度が 76.7% と 56.7% となっており，精度がかなり上昇している．一方で，「レーシック手術は痛い」のトピックの結果において，根拠関係のリンクの重みが 0 から 10 の間では，正解パッセージの割合よりも低い結果となっている．これは，TextRank が頻出語を多く含むパッセージに高い重要度を与える傾向にあり，「レーシック手術は痛い」のトピックの正解パッセージには，もう一方のトピックに比べて頻出語が少なかったためである．しかしながら，根拠関係のリンクの重みを増やすことで，頻出語が含まれていない重要なパッセージも抽出することができている．

今後の予定は，根拠関係のリンクの重みの最適値について検討することである．図 4にあるように，リンクの重みを 50 にしたときに一番高い精度にはなっていない．これは，根拠関係のみがパッセージが重要かどうかを決定づけているわけではないということを意味している．それゆえに，根拠以外の他の意味関係のリンクを導入することも今後の予定である．

5 まとめ

本稿では，単語，文，パッセージ，文書の各層に分かれた多層ネットワーク構造を導入し，HITS アルゴリズムを用いて層間の包含関係と文ノード間の根拠関係を考慮した重要パッセージ抽出手法を提案した．まず，一般的な目的の要約と情報信憑性判断のための要約との差異について検討し，情報信憑性判断のためには重要パッセージ抽出を中心とした要約が有効であると考えた．我々はサーベイレポートを生成するシステムの開発を試みており，そのシステムにおける要素技術として TextRank に基づく重要パッセージ抽出手法を検討した．我々は TextRank に意味的関係を考慮させることで精度の上昇が出来ると考えた．各層間の包含関係を考慮すること，サーベイレポート作成に必要な部品を容易に作成できるようになることから，多層ネットワーク構造を導入した．多層ネットワーク構造に文ノード間の根拠関係と層間の包含関係を導入し HITS アルゴリズムに反映させ，TextRank による根拠関係を考慮した重要パッセージ抽出を行う手法を提案した．評価実験を行った結果，提案手法を用いると，上位のパッセージの抽出精度が高くなる事を確認した．今後，根拠関係のリンクの重みの最適値について検討をすることを予定している．また，根拠以外の他の意味関係のリンクを導入することを検討している．

謝辞

本研究は，独立行政法人情報通信研究機構の委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」プロジェクトの成果である．

参考文献

[1] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, Vol. 30(1-7), pp. 107–117, 1998.

[2] Koichi Kaneko, Hideyuki Shibuki, Masahiro Nakano, Rintaro Miyazaki, Madoka Ishioroshi, and Tatsunori Mori. Mediatory Summary Generation: Summary-Passage Extraction for Information Credibility on the Web. In *the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 23)*, pp. 240–249, 2009.

[3] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into texts. In *The 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP2004)*, pp. 404–411, 2004.

[4] Koji Murakami, Eric Nichols, Suguru Matsuyoshi, Asuka Sumida, Shouko Masuda, Kentaro Inui, and Yuji Matsumoto. Statement Map: Assisting Information Credibility Analysis by Visualizing Arguments. In *the 3rd Workshop on Information Credibility on the Web (WICOW2009)*, pp. 43–50, 2009.

[5] Dragomir R.Radev. A common theory of information fusion from multiple text sources step one: cross-document structure. In *the 1st SIGdial workshop on Discourse and dialogue*, pp. 74–83, 2000.

[6] Mann William and Sandra Thompson. Rhetorical structure theory: towards a functional theory of text organization. In *Text*, Vol. 8, pp. 243–281, 1988.

[7] 村上浩司, 増田祥子, 松吉俊, 乾健太郎, 松本裕治. 言明間の意味的関係の体系化とコーパス構築. 言語処理学会 15 回年次大会発表論文集 pp.602–605, 言語処理学会, 2009.

[8] 富田紘平, 高村大也, 奥村学. 重要文抽出と文圧縮を組み合わせた新たな抽出的要約手法. 自然言語処理研究会報告 2009-NL-189, pp.13–20, 情報処理学会, 2009.