Webを利用した語の関連度の視覚化による文章の信憑性判断支援

中林　猛† 湯本　高行†† 新居　学†† 高橋　豊†† 角谷　和俊†††

† 兵庫県立大学大学院工学研究科　〒 671-2201 兵庫県姫路市書写 2167
†† 兵庫県立大学大学院工学研究科　〒 671-2201 兵庫県姫路市書写 2167
††† 兵庫県立大学環境人間学部　〒 670-0092 兵庫県姫路市新在家本町 1 丁目 1-12
E-mail: †er10j040@steng.u-hyogo.ac.jp, ††{yumoto,nii,takahasi}@eng.u-hyogo.ac.jp, †††sumiya@shse.u-hyogo.ac.jp

あらまし　ユーザは Web で特定の情報を取得する際，複数のページを見比べてその情報の信憑性を判断する．しかし，その取捨選択はユーザにとって負担のかかる行為であり，その事柄について詳しくない者にとっては判断すら困難である．よって我々は，ユーザが自ら情報の真偽を判断するための材料を提示する事で支援する方法を提案する．提示された材料を利用して文章信憑性を判断する事で，ユーザの負担を軽減できると考える．この目的達成のために，我々は主題とそれを説明するために用いられている語の関連度を視覚的に提示する方法が有効であると考えた．これは，主題と関係の強い語を用いている文章は典型的であり，信用に足るという考えに基づく．語間の関係の強さは Web 検索エンジンを用いて算出する．その関連度は文書と同時にユーザに提示する．そのために，ユーザインタフェースの提案も行う．予備実験では，32 種類の文章を入力として，主題を表現する語の抽出と，主題とそれを説明する語の関連度の算出をそれぞれ独立に行った．その結果，語の抽出実験では約 4 割の精度で抽出できた事を確認した．しかし，正しく抽出できないいくつかの問題が確認できた．また，関連度の算出実験ではほぼ期待される関連度が算出された事を確認した．しかし，いくつかの文章では期待される関連度と異なる数値が算出される問題も確認できた．

キーワード　語の関係性，情報信憑性，特徴語抽出

Support for Credibility Judgment of Information by Visualization of Relation between Words on the Web

Takeru NAKABAYASHI†, Takayuki YUTAKA††, Manabu NII††, Yutaka TAKAHASHI††, and Kazutoshi SUMIYA†††

† Graduate School of Engineering, University of Hyogo 2167 Shosha, Himeji, Hyogo, 671-2280 Japan
†† Graduate School of Engineering, University of Hyogo 2167 Shosha, Himeji, Hyogo, 671-2280 Japan
††† School of Human Science and Environment, University of Hyogo 1-1-12 Shinzaike-honcho , Himeji, Hyogo, 670-0092 Japan
E-mail: †er10j040@steng.u-hyogo.ac.jp, ††{yumoto,nii,takahasi}@eng.u-hyogo.ac.jp, †††sumiya@shse.u-hyogo.ac.jp

1. はじめに

近年 Web における情報量は膨大なものとなっている．その中から，適切に情報を取捨選択することは，知識の無い者にとって正しい判断が困難である．その理由として，情報量が膨大であること，偽の情報が数多く存在することにある．このような情報の真偽をユーザが判断する場合，自ら検索エンジンを利用して複数のページを見比べ，その信憑性を判断する．しかし，この作業はユーザにとって非常に負担が大きい．Web 上に存在する実際の例を図 1 に示す．これは，鉛筆について記述されている，Wikipedia（左）と Uncyclopedia（右）の記事である．多くの者は鉛筆についての知識があるので，Uncyclopedia の記事は偽情報だと容易に判断できる．しかしながら，知識の無い事柄について書かれていた場合，正しい判断は困難である．ここで，両文章を注意深く見ると，Wikipedia の文章は Uncyclopedia の文章に比べ典型性が高い (一般的である) ことがわかる．

そこで本研究では，ユーザ支援のために文章の典型性に着目した材料の提示を目的とする．本研究では，“当たり障りのない文章は信用に足る” と考える．つまり，偽の情報発見の逆問題と捉え，文章の典型性を信憑性尺度の 1 つとした．我々が考

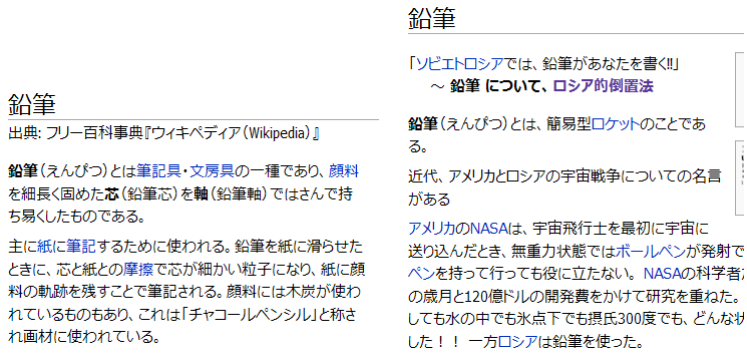

図 1　Web 上の文章例

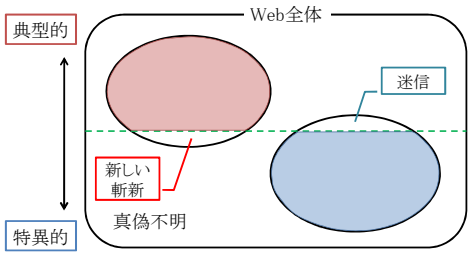

図 2　真偽情報と文章典型性の関係

えている真偽情報と文章典型性の関係を図 2 に示す．図に示すように，真情報の大部分が典型的な文章であり，偽情報の殆どが特異的な文章と考えている．また，典型的な偽情報の例として “迷信” や “有名なネットスラング” 等があげられる．さらに，典型的でない情報の例としては “新しい情報” や “斬新なアイデアの文章” 等があげられる．加えて，真偽が存在しないような情報もある．これに区分される文章は，宗教のような立場で真偽や典型性が異なる文章や未だ解明されていない事実である．本研究が対象とする情報は図中の真情報と偽情報等がある．

我々は，文章の典型性は主題と文章中の語の関連度で表すことができると考えた．そこで本研究では，文章の典型性を表す “主題と文章中の語の関連度” を提示することで，ユーザの支援を行う．提示される材料は “直感的に理解できる” ことが望ましい．よって主題と文章中の関連度を視覚化し提示する必要がある．これを実現したユーザインタフェースのイメージを図 3 に示す．

図 3　ユーザインタフェースイメージ

図の左側が閲覧している Web ページである．システムは疑わしい文章を強調しユーザに提示する．図の右側がシステムの算出した主題に対する各語の関連度をグラフ化したものである．これはシステムが疑わしい文章を強調した根拠として，ユーザに提示する．

このシステムを構築する提案手法は大きく分けて，主題を表現する語の抽出ステップと，主題とその主題を説明するために用いられている語の関連度算出ステップの 2 つの手法から構成する．これら 2 つの手法で得られた最終的な結果を図 3 の様に視覚化し提示する．主題を表現する語の抽出では，“主語であるか”，“Web ページタイトルに使用され易いか” という 2 点に着目している．また，語間の関連度算出には Web 検索エンジンを用いた語の共起頻度を用いる．

次章より，システムを構築する提案手法の詳細を述べる．

2.　テーマ語と説明語の抽出および関連度算出

本研究では Web 上のテキストでも，いわゆる Wikipedia 記事のような “ある事柄，事実について説明記述がなされた文章” を対象として，大きな 2 つのプロセスにより解析する．ユーザに解析結果を提示するまでの大まかな処理の流れを以下に示すと共に，図 4 に処理の概要図を示す．

（1）　解析対象の Web ページの文章から “テーマ語”，“説明語” の抽出

（2）　検索エンジンを用いて，“テーマ語” と “各説明語” の関連度を算出し，その関連度を棒グラフとしてユーザに提示

図 4　処理の流れ

なお，文章から語を抽出する際に，形態素解析器である茶筌を用いる [1]．次節より，それぞれのプロセスについて詳しく述

べる．

2.1 テーマ語と説明語の抽出

本研究では，任意の文章の主題を表現している語を “テーマ語”，主題を説明するために用いている語を “説明語” と定義する．

テーマ語はその特徴として，以下の 2 つがあげられる．

（1） 主語である事が多い

（2） Web ページのタイトルに用いられる事が多い

したがって，この 2 つの特徴を満たす語をテーマ語として抽出する．まず，(1) の特徴から，「“は” や “とは” の前に来る名詞」をテーマ語の候補語 $c \in C_p$ とする．C_p はページ p におけるテーマ語候補語集合である．次に，(2) の特徴から，(1) 式を用いてテーマ語を決定する．

$$t = \arg\max_{c \in C_p} \left(\frac{\mathrm{Search}(intitle(c) \wedge e_i)}{\mathrm{Search}(c \wedge e_i)} \right) \quad (1)$$

ただし，$\mathrm{Search}(c \wedge e_i)$ は候補語 c と文章中に i 番目に出現する説明語 e の And 検索結果件数であり，$\mathrm{Search}(intitle(c) \wedge e_i)$ はタイトルに候補語 c が含まれかつ，i 番目の説明語 e が含まれる Web ページの検索結果件数である．

また，テーマ語以外の名詞は Web ページの著者がテーマについて説明するために，意図して使用した単語であると考えられる．したがって，テーマ語以外の名詞を説明語として抽出する．

2.2 テーマ語と各説明語の関連度算出

語間の関連度を表す指標は様々なものがある．特に Web 上での語の共起頻度を表す指標は，Jaccard 係数や Simpson 係数，WebPMI などが主に知られている．本研究では，主題とその説明に用いられる語の関連度を算出する．このことから，テーマ語と説明語の出現ページ集合が図 5 に示すような包含関係になりやすい事が予測できる．これは，テーマ語の “鉛筆” に対し，“鉛筆削り”，“筆記具” の 2 語が説明語として用いられた例である．説明語の “鉛筆削り” は，“鉛筆” というテーマを説明する場合のみに用いられる語であるが，“筆記具” は他のテーマでも使用され，テーマ語である “鉛筆” を包含している．このような場合，説明語の “鉛筆削り”，“筆記具” はテーマ語の “鉛筆” と関連度が高く算出されるべきである．このように双方の観点で関連度が高く算出されるべきであるから，ここでは Simpson 係数をそのまま用い，語の関連度 Rel とする．語 A と語 B の Simpson 係数は式 (2) で定義される．

$$\mathrm{Rel}(A, B) = \mathrm{Simpson}(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)} \quad (2)$$

要素数の少ない語を基準として語間の関連度を算出する Simpson 係数では，上記で述べた包含関係でも関連度が低く算出されることがなく，本手法に適していると言える．

図 5 は，それぞれの語が含まれた Web ページ集合を表している．次章より，これらの手法を実装したユーザインタフェースについて詳しく述べる．

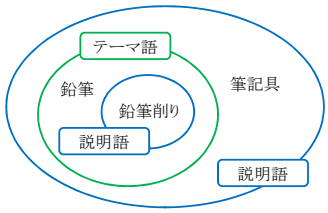

図 5　Web ページ集合の包含関係

3. 信憑性判断支援のための語の関連度の視覚化

我々は，情報信憑性判断支援には文章の注意すべき部分の提示と，それを裏付ける根拠の提示が必要であると考えている．これらを実現したユーザインタフェースのイメージを図 3 に示した．

図の左側が閲覧している Web ページの記事であり，右側がシステムの算出したテーマ語に対する各説明語の関連度をグラフ化したものである．まず，このシステムの使用方法を説明する．ユーザが Web ページを閲覧している最中に解析したい文章を見つけた場合，解析ボタンをクリックする．もしくは，特定範囲の文章を解析したければ文章を選択しグラフ領域内にドラッグ＆ドロップする．すると，システムは注意すべき文章を強調し，その根拠としてテーマ語と説明語の関連度をグラフ化してユーザに提示する．

次に，システムによる文章の強調の意味と強調方法について述べる．強調色が濃いほど注して読むべき部分である．強調色は，テーマ語と説明語の関連度 Rel に対応している．以下に強調条件を示す．

- $Rel < 0.2$：Orange
- $0.2 \leqq Rel < 0.3$：Wheat
- $0.3 \leqq Rel < 0.4$：Oldlace

これにより，ユーザは疑わしい文章を一目で把握する事ができる．また，強調の範囲は，関連度を算出した説明語が含まれる一文である．なお，一文中に関連度を算出した説明語が 2 語以上含まれる場合，関連度の最も低い説明語を優先し，文を強調する．

最後に，システム関連度のグラフ化について述べる．2.2 節で述べた，テーマ語と説明語の Simpson 係数をそのままグラフ化して提示する．これにより，ユーザは注意する根拠を理解でき，次回からの閲覧にこの経験を活かす事ができると期待される．なお，解析ボタンをクリックした場合，システムは Web ページの本文を抽出する．この際，本文抽出プログラム ExtractContent を利用している(注1)．

4. 予 備 実 験

2. 章で述べた手法を用い，予備実験を行った．今回は 2.1 節のテーマ語と説明語の抽出および 2.2 節の各語の関連度算出の 2 つの実験を個別に行い，それぞれが有効な手法であるか検討した．

(注1)：http://search.cpan.org/dist/HTML-ExtractContent/

以下に示す表 1 に含まれるタイトルの記事を Wikipedia と Uncyclopedia からそれぞれの概要部分を収集し，入力文章とした．

表 1　入力テキスト記事タイトル

オブジェクト指向	ウエイトトレーニング	原子力
日本放送教会	コーラ	赤木しげる
Hyde	ドラゴンボール	地球温暖化
Windows	ノストラダムス	天動説
アメリカ合衆国	ピザ	
イチロー	ペン回し	

4.1　テーマ語抽出実験

2.1 節で述べた手法を用い，表 1 の入力文章からテーマ語を抽出した．システムが抽出した結果を表 2 に示す．

表 2　テーマ語抽出結果

記事タイトル	Wikipedia	Uncyclopedia
Hyde	Hyde	Hyde
NHK	運用資金	日本放送協会
Windows	マイクロソフト	屋外
アメリカ合衆国	合衆国	アメリカ軍
イチロー	日本時代	仰木彬
ウエイトトレーニング	ウエイトトレーニング	
オブジェクト指向	オブジェクト指向	オブジェクト指向言語
コーラ	コーラエキス	種子
ドラゴンボール	正式表記	鳥山
ノストラダムス	ノストラダムス	ノストラダムス
ピザ	ピッツァ	ためイタリア人
ペン回し	ペン回し	ペン回し
原子力	原子核変換	放射性物質
地球温暖化	地球温暖化	日本政府
天動説	天動説	天動説
赤木しげる		アカギ

また，抽出結果の精度を表 3 に示す．

表 3　テーマ語抽出精度

	正解語数	精度
記事タイトルと同文字列	11	0.34
意味的に正解	15	0.47
平均	13	0.41

実験結果から，平均して約 4 割の記事において，正しくテーマ語が抽出できたことがわかる．しかし，“イチロー” の記事のように正解の語が抽出できなかった例や，“ウエイトトレーニング” の記事のように，何も抽出されなかった例も確認できた．これは，人名などは正しい表記ではなく愛称で表記される場合があるため，Web ページのタイトルにあまり使用されていなかったことが原因としてあげられる．また，“〜〜とは” の形式で記述されていない文章の場合，現在の手法ではテーマ語候補となる語が検出されないため，何も抽出されない結果になったと考えられる．

また，抽出結果の精度から，記事タイトルと同文字列の語を抽出する事は難しいが，意味的にテーマ語として正しい語を抽出する事は，改善次第で可能であると期待できる．具体的な改善法としては，2.1 節で述べたテーマ語の候補を決定する条件の緩和を考えている．また，いくつか研究されている，主題自動抽出法(注2)の利用も考えている．

4.2　テーマ語と各説明語の関連度算出および視覚化

2.2 節で述べた手法を用い，表 1 の入力文章から各語の関連度を算出し，棒グラフで表した．ただし，テーマ語は人手で抽出したものを用いた．システムが算出した結果の一部を図 6，7 に示す．

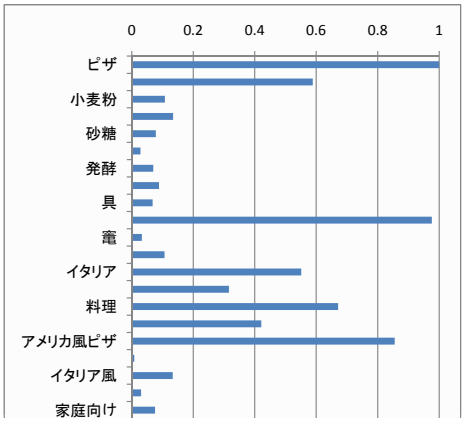

図 6　ピザ：Wikipedia

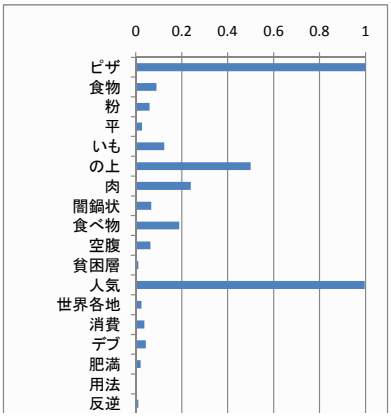

図 7　ピザ：Uncyclopedia

語の順番は，文章に出現した順番である．図を見比べると，Uncyclopedia 記事から生成されたグラフに比べ，Wikipedia 記事から生成されたグラフの方が，若干高い関連度を示す語が多い事がわかる．しかし，図 6 の “小麦粉” や，図 7 の “粉” のように，テーマ語と関連が強いと思われる語の関連度が低く示された例も確認できた．これは，一般的には周知の語であったために，関連度が低く算出された事が原因と考えられる．また，Web ページの著者がテーマについて過度に詳しい場合，一般的な記事ではされる事の無い説明を記述する場合がある．この場合も，実際は関連が強い説明語であっても低い値を算出する．

(注2)：野本 忠司：確率モデルによる主題の自動抽出 他

加えて，図 7 の “の上” のような形態素解析器による，本来意図しない語の抽出も見られた．

5. 関 連 研 究

5.1 情報信憑性に関する研究

Zoltán らは TrustRank と呼ばれる概念を発表している [2]．この概念はスパムページの判定に用いることができるとしている．TrustRank とは，あらかじめ “信頼できる” と人手で判断された良質なページを拠点とし，リンク先にスコアを付ける．ここで，拠点ページから直接リンクされているページには多くのスコアを与え，離れるほど低いスコアを与える．一般的に，良質なページは滅多に悪質なページにリンクを張らず，逆に悪質なページは悪質なページにリンクを張ることが多い．したがって，TrustRank の概念を用いることで良質なページとスパムページに分類することができ，良質なページのみを取得する事が可能である．以上のように，TrustRank は周囲のページから信憑性を推定するアプローチである．本研究では，TrustRank とは異なり “文章自体” から信憑性を推定するというアプローチである．

また，山本らは，検索結果の集約とページ生成時間分布解析を用いた，Web 情報の信用度評価システム “ほんと？サーチ” を開発している [3]．このシステムは，検索結果の集約と Web ページ生成時間分布の解析により，精度の高い信用度評価を可能としている．“ほんと？サーチ” は，語のフレーズを用いて信用度の算出と評価を行っている．そのシステムの特性上，信用度評価を行う対象が “文（一文）” であり，長い “文章” には不適と言える．本研究では，“文章” を評価の対象としている．

5.2 文章の典型性に関する研究

我々は文章の典型性を語の関係から推定できると考えている．語の関係を判定する手法の 1 つとして，小山らは Web ページの文書構造を利用した関連キーワード抽出法と題して，詳細語抽出アルゴリズムを提案している [4]．これは，あるキーワードを詳細化する語（詳細に説明するため用いられる様な語）を抽出する手法である．以前我々は，この手法の詳細語判定の部分を用いて文章の典型度を算出する手法を提案した [5]．しかしながら，文章に 1 つの典型度を算出し，数値をそのままユーザに提示する手法は直感的に把握しづらい．加えて，文章の一部分のみ偽の情報が書かれていた場合，典型度を正確に数値で算出する事が困難であることが分かった．

そこで本研究では，ユーザが直感的に注意して読むべき部分を把握できるシステムが必要だと考えた．よって本稿では，主題と各語の関連度を視覚化して提示する手法を提案した．本手法で，部分的に疑わしい文章も一目で把握できると期待される．

6. お わ り に

本稿では，文章の信憑性判断支援を目的として，文章典型性の視覚化手法を提案した．手法は 2 つのステップから構成され，テーマ語の抽出法と各語の関連度算出法を考案し，それぞれ独立に予備実験を行った．予備実験結果から，2 つの手法それぞれにいくつかの問題があることが分かった．テーマ語の抽出法では，正しくテーマ語が抽出されない問題と，何も抽出されない問題が確認できた．これらは，テーマ語が正式表記でない事や，テーマ語を抽出する条件に当てはまる語が無い事が原因と考えられる．また，各語の関連度算出法では，本来関連の強いであろう語が低く示される問題と，形態素解析器による意図しない語の抽出が確認できた．これらは，説明語が一般的過ぎる語であることや，著者が過度に詳細な説明を記述した事が原因として考えられる．加えて，抽出した説明語が多過ぎるため，視覚化された関連度を有効的にユーザに提示し難い問題も確認できた．今後，これらの問題を解決するために手法の改善を行う予定である．具体的には，テーマ語抽出の条件を緩和することで良い結果が期待できる．もしくは，テーマ語の抽出に “主題の自動抽出法” 等を利用する事も考えている．関連度の算出には，Web の検索結果件数を用いず，Wikipedia シソーラスを用いる手も考えている．また，用いる説明語をある程度選択し，グラフを生成することでより良い結果が期待できると考えている．

謝　　辞

本研究の一部は，平成 22 年度科研費基盤研究 (B)(2)「ユーザの潜在的意図を用いたレス・コンシャス情報検索基盤の構築」(課題番号：20300039) によるものです．ここに記して謝意を表すものとします．

文　　献

[1] “茶筅”. http://chasen-legacy.sourceforge.jp/.

[2] J. P. Zoltán Gyöngyi, Hector Garcia-Molina: “Combating web spam with trustrank”, Proceedings of the Thirtieth international conference on Very large data bases, **30**, (2004).

[3] 山本祐輔, 手塚太郎, アダムヤトフト, 田中克己：“ほんと？サーチ：検索結果の集約とページ生成時間分布解析による web 情報の信用度評価”, 日本データベース学会 Letters, **6**, 1, pp. 53–56 (2007).

[4] S. Oyama and K. Tanaka: “Query modification by discovering topics from web page structures”, Proceedings of the Sixth Asia Pacific Web Conference (APWEB’04) (2004).

[5] T. Nakabayashi, T. Yumoto, M. Nii, Y. Takahashi and K. Sumiya: “Measuring peculiarity of text using relation between words on the web”, Proceedings of the role of digital libraries in a time of global change, and 12th international conference on Asia-Pacific digital libraries, 4, pp. 112–115 (2010).