様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 ２９ 日現在

機関番号：２４５０６
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００４２９

研究課題名（和文）Webページの大域的・局所的特徴の可視化による情報信頼性の判断支援方式の研究

研究課題名（英文）A Study on Supporting Information Reliability Judgment by Presenting Global and Local Features of Web Pages

研究代表者
湯本　高行（Yumoto, Takayuki）

兵庫県立大学・工学研究科・助教

研究者番号：２０４５３１５２

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：本研究では，ユーザがWeb上の情報の信頼性を判断する材料として，文の典型性と文体に注目し，それぞれを推定してユーザに提示する手法を開発した．文の典型性推定では，与えられたキーワードに対する文の典型性を，語の共起関係に基づいて推定する．文体の推定では，文体クラスとして，敬体，常体，会話体，俗語体を定義し，文末の表現に注目してこれらの文体への分類を行う．文体分類には，文体ごとに異なる情報源から自動構築した文末表現辞書を用いる．

研究成果の学術的意義や社会的意義
本研究では，文の典型性の推定手法と文体分類手法を開発した．文の典型性推定では，語の共起確率だけでなく，その予測値を併用することで推定精度を向上させた．文の典型性の推定結果をユーザに提示することにより，ユーザは典型性の低い文章については世間一般に認知されていない情報として警戒して閲覧することができるようになる．また，文体分類においては，定義した4つの文体に対して高い精度での分類を実現した．さらに，文体分類の応用として，文体による注目トピックの違いや文体と文章の難易度の関係についての分析を行った．

研究成果の概要（英文）：We proposed methods to estimate the typicality and writing style of sentences. They are used as criteria to judge the reliability of Web pages. The sentence typicality for a given keyword is estimated based on the co-occurrence relationship between words. In the writing style classification, four writing style classes are defined, distal style, direct style, conversational style, and slang style. We focus on sentence-final expression for the writing style classification, and we use sentence-final dictionaries automatically constructed from information source different for each style.

研究分野：情報検索

キーワード：情報信頼性　Webマイニング　典型性　文体分類

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

Web 上の情報は一般のユーザにも身近であり，その重要性は増している．これを裏付ける調査としては，総務省 情報通信政策研究所が実施したアンケート結果をまとめた，情報通信メディアの利用時間と情報行動に関する調査[1]がある．この調査では，Web 上の情報 (報告書内では「インターネット」と記載) は，テレビ (91.0%) に次いで情報源として重要 (69.8%) であるとの回答が得られており，これは新聞 (67.7%) を上回っている．その一方で，信頼度はテレビ (62.7%)，新聞 (68.6%) に比べて著しく低い，29.7% であった．すなわち， Web 上の情報はその重要性にも関わらず， 依然として信頼性が低い状況が続いており，情報が信頼できるか否かをユーザが判断することを支援する仕組みが必要であると言える．

２．研究の目的

Web 上の情報の信頼性の判断基準としては，ページ単体を分析して得られる局所的な特徴と他のページと比較して得られる大域的な特徴の２つに分けられる．本研究では，局所的な特徴としては文体に，大域的な特徴としては典型性に注目し，それぞれを推定してユーザに提示する手法を開発することを目的とする．

３．研究の方法

本研究では，Web ページの大域的特徴として，(1)文の典型性に注目し，その推定手法の開発を行う．また，局所的特徴としては，(2)文章の文体に注目し，文体分類の手法の開発を行う．

(1)文の典型性推定では，与えられたキーワード（クエリ）に対する文の典型性を，文を構成する語のクエリへの関連度の平均値を用いるため，クエリに対する語の関連度の推定手法を開発する．語の関連度は語の共起関係に基づいて算出する．

(2)文章の文体の推定では，文章の丁寧さの観点から文体クラスとして，敬体，常体，会話体，俗語体の 4 クラスを定義し，文章をこれらに分類する手法を開発する．日本語の場合，文体の特徴は文末に現れやすいと考え，文体ごとに異なる情報源を用いて文末表現辞書を構築し，これらの辞書を用いて文体の分類を行う．

４．研究成果

(1) 文の典型性推定

文の典型性推定には文を構成する語の関連度の平均値を用いるが，語の関連度としてコーパス中での語の共起確率を用いる手法を開発した．この手法では，語の関連度を，語が出現する文書のうちクエリも出現する割合とした．さらに，同義語を考慮して，語の出現文書数および共起回数を補正した値を関連度の算出に用いた．この手法について評価実験を行い，語が出現する文書数および語の共起回数の，関連度に対する影響について分析したところ，上記の関連度の算出においては語が出現する文書数の値が支配的であり，語の共起回数の影響が非常に小さいことが明らかになった．

このことを踏まえ，コーパス中での語の出現文書数から予測される共起確率に対する実際の共起確率の比を語の関連度とする手法を開発した．この手法では，測定値より実際の値が大きいほど関連が強いと考える．共起確率の予測値の算出のイメージを図 1 に示す．まず，あるクエリと共起するすべての名詞について出現文書数と共起確率の散布図を作成する．続いて，出現文書数の値に応じて分割した各区間において共起確率の平均値を算出する．各区間の中央に平均値が存在するとみなし，これらの平均値をつないだ折れ線によって予測値を算出する．

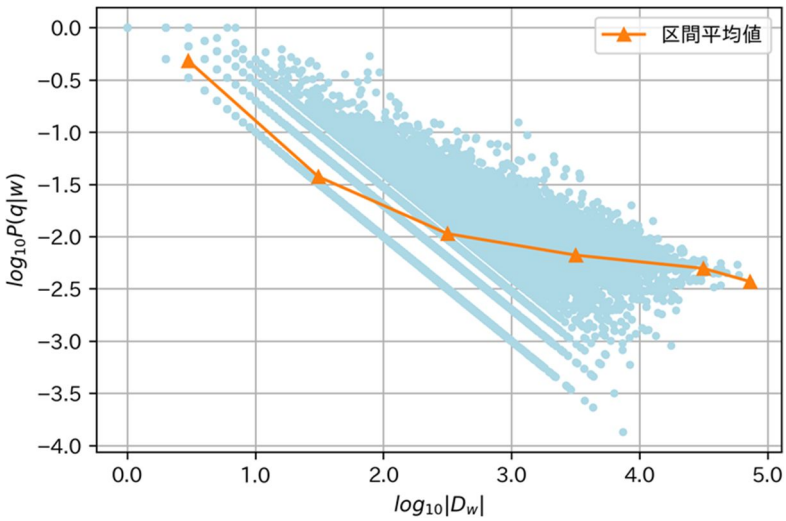

図 1 対数共起確率の予測値の算出イメージ
（縦軸が語の対数共起確率，横軸が語の出現文書数の対数）

語の関連度についての評価実験では，20 種のクエリに対して 5 文ずつの計 100 文に対して，クラウドソーシングによって関連の深さを４段階で採点させ，語の関連度と評価者の点数との相関係数により評価を行った．その結果，語の共起確率を関連度として用いた場合に比べ，語の共起確率の予測値を用いた手法では 20 クエリ中 15 クエリで相関係数が向上し，有意水準 1％で有意な差が見られた．

また，文の典型性については，上記に加えて，類似する高頻度の語への置き換えなどによる改良を行った．これは，出現頻度の低い語については語の共起数も少なくなるため，共起に基づいて語の関連性を把握することが困難であるため，類似する高頻度の語に置き換えたうえで典型性を推定するものである．語の類似度の算出には，日本語 Wikipedia エンティティベクトルのコサイン類似度を用いた．これにより，低頻度の語においては文の典型性の推定精度を向上させることができた．

(2) 文体分類

文章の丁寧さの観点から以下の 4 つの文体を定義し，各文の文末に注目して分類を行う．

- 敬体：ですます調
- 常体：だ・である調
- 会話体：敬体・常体の後に終助詞「ね，よ，な，か，かしら，の」が続く文章．たとえば，「～ですね」などが該当する．
- 俗語体：ネットスラングなどを含む砕けた文章

文末は，各文の形態素解析結果の末尾から動詞，形容詞，名詞のいずれかまでと定義する．文章の文体を分類するため，俗語体を除く 3 つのクラスで文末表現辞書を構築する．この文末表現辞書には文体ごとに異なる情報源から抽出した文末表現を用いる．具体的には，敬体はニュースサイトの記事を，常体は新聞記事を，会話体は Yahoo!知恵袋の質問文および回答文を使用する．ここで，常体の情報源として用いる新聞記事には会話文やコラムが存在するため「～です」や「～だね」のような常体クラスに属さない表現が抽出されることがある．そこで，常体辞書内から敬体辞書や会話体辞書にも共通して見られる表現を取り除く．さらに，敬体辞書と会話体辞書で取り除くことができなかった表現への対策として，常体辞書に収録する表現にしきい値を設け，出現回数がしきい値以上となった表現のみを辞書に用いる．同様に，会話体の情報源として用いた Yahoo!知恵袋の文章においても会話体以外の表現が含まれているため，会話体辞書についても予備実験によって求めたしきい値を用いる．

文体分類では，まず，各文から文末表現を抽出し，文末表現を各辞書と照らし合わせ，文体を決定する．文体の決定順は，辞書に含まれているノイズが少ない敬体辞書，常体辞書，会話体辞書の順に照合し，いずれかの辞書の表現と完全一致した場合に，その文の文体を辞書の文体と一致させる．文章全体の分類では，各文の文体を用いて，以下の順にルールを適用して文体を決定する．

- 文章中で俗語体の文の割合がしきい値以上であれば文章全体を俗語体に分類する．
- 文章中の会話体の文の割合がしきい値以上であれば文章全体を会話体に分類する．
- 上記以外で文章中の敬体表現と常体表現の文の割合を比べたとき，敬体表現の方が多ければ文章全体を敬体に，常体表現の方が多ければ文章全体を常体に分類する．同数のとき，敬体と常体の両方に分類する．

評価実験では，正解データとして，敬体，常体，会話体については辞書構築に用いたサイトから辞書構築に用いていない文章を，俗語体については，はてなブックマークのコメント欄を収集して用いた．提案手法による分類の F 値は，敬体が 0.976，常体が 0.953，会話体が 0.867，俗語体が 0.853 となった．

また，文体分類の応用として，ある話題についてのツイートで言及されやすいトピックが文体によって違うかを分析した．この実験では，トピックは同一の話題について異なる文体を含むツイート集合に対して LDA[2]を適用することで求めた．その結果，話題によるが，言及されやすいトピックに明確に違いがある場合が確認された．また，文体によって文章の難易度が違うかを文章難易度判断システムの jReadability[3]を用いて求めたところ，難易度のスコアは難易度の高い順に常体，敬体，会話体，俗語体であり，簡易的ではあるが，文体によって文章の難易度を推定することができることを確認した．

参考文献

[1] 総務省情報通信政策研究所，平成 27 年情報通信メディアの利用時間と情報行動に関する調査報告書，2016．

[2] D. M. Blei, A. Y. Ng, M. I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol.3, pp.993-1022, 2003.

[3] Y. Hasebe, J. H. Lee. Introducing a Readability Evaluation System for Japanese Language Education. Proc. of the 6th International Conference on Computer Assisted Systems for Teaching & Learning Japanese (CASTEL/J), pp.19- 22, 2015.

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 飯塚 翔、湯本 高行、新居 学、上浦 尚武	４．巻 J101-D
２．論文標題 含意関係に基づく二部グラフを用いた情報の断片のランキング	５．発行年 2018年
３．雑誌名 電子情報通信学会論文誌D 情報・システム	６．最初と最後の頁 681～689
掲載論文のDOI（デジタルオブジェクト識別子） 10.14923/transinfj.2017DEP0008	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 T. Yumoto, T. Yamanaka, M. Nii, and N. Kamiura	４．巻 22
２．論文標題 Finding Rare Information from the Web Using Social Bookmarks and Word Co-occurrence	５．発行年 2017年
３．雑誌名 International Journal of Biomedical Soft Computing and Human Sciences	６．最初と最後の頁 9-18
掲載論文のDOI（デジタルオブジェクト識別子） 10.24466/ijbschs.22.1_9	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計11件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 村本 直樹、大島 裕明、湯本 高行
２．発表標題 語の係り受け関係と分散表現を用いたレビューからの属性と意見の抽出
３．学会等名 第9回ソーシャルコンピューティングシンポジウム
４．発表年 2018年

１．発表者名 小山 雄也、湯本 高行、礒川 悌次郎、上浦 尚武
２．発表標題 語の共起の実測値と予測値に基づく 名詞の組の関連性推定
３．学会等名 第11回Webとデータベースに関するフォーラム
４．発表年 2018年

１．発表者名
新本 拓也、湯本 高行、金子 周司、礒川 悌次郎、上浦 尚武

２．発表標題
身体部位の表現の違いを考慮したQAサイトからの病訴の検索

３．学会等名
第11回Webとデータベースに関するフォーラム

４．発表年
2018年

１．発表者名
Yuya Koyama, Takayuki Yumoto, Teijiro Isokawa, Naotake Kamiura

２．発表標題
Measuring Term Relevancy based on Actual and Predicted Co-occurrence

３．学会等名
13th International Conference on Ubiquitous Information Management and Communication（国際学会）

４．発表年
2019年

１．発表者名
小山雄也，湯本高行， 礒川悌次郎， 上浦尚武

２．発表標題
類義語を考慮した自己相互情報量に基づく文単位の典型性推定

３．学会等名
第10回Webとデータベースに関するフォーラム

４．発表年
2017年

１．発表者名
有馬直也，湯本高行， 礒川悌次郎， 上浦尚武

２．発表標題
文体と意見極性に基づくツイートの分類

３．学会等名
第10回Webとデータベースに関するフォーラム

４．発表年
2017年

１．発表者名
有馬直也，湯本高行，　礒川悌次郎，　上浦尚武

２．発表標題
文末表現辞書を用いた文体分類とその応用

３．学会等名
第10回データ工学と情報マネジメントに関するフォーラム

４．発表年
2018年

１．発表者名
中谷将佳史，湯本高行，礒川悌次郎，上浦尚武

２．発表標題
Wikipediaのカテゴリ情報を用いたツイート発信者の特徴表現

３．学会等名
情報処理学会第169回データベースシステム研究発表会

４．発表年
2019年

１．発表者名
吉田恭輔，湯本高行，金子周司，礒川悌次郎，上浦尚武

２．発表標題
構文木と専門用語辞書を用いた医学論文からの未知用語の発見

３．学会等名
情報処理学会第169回データベースシステム研究発表会

４．発表年
2019年

１．発表者名
新本拓也，湯本高行，金子周司，礒川悌次郎，松井伸之，上浦尚武

２．発表標題
QAサイトでの共起に基づく患者の自覚症状入力支援

３．学会等名
情報処理学会第170回データベースシステム研究発表会

４．発表年
2019年

１．発表者名 服部雄也，湯本高行,芦田真一,井上直樹，礒川悌次郎，上浦尚武
２．発表標題 類似工場推薦のための特徴表現
３．学会等名 第12回データ工学と情報マネジメントに関するフォーラム
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考