Vol.2011-DBS-153 No.5
2011/11/3

語の連想関係に基づく文章の典型性の可視化

中林 猛[†1] 湯本高行[†1]
新居 学[†1] 高橋 豊[†1]

近年，Web で情報収集を行うことが多いが，情報の取捨選択に，文章が典型的か否かで判断することが多々ある．そこで我々は，文章の “典型的でない記述の箇所” と “その根拠または補助となる情報” を視覚的に提示する手法を提案する．提案手法は，語の連想関係を Web を用いて抽出し，抽出された連想関係に基づきグラフを構築する手法と，構築されたグラフの形状による分類を行う手法の 2 段階で文の典型性を判定する．また，構築したグラフを典型性の根拠として提示する．実験では，それぞれの手法に用いる数式のしきい値決定実験と，決定されたしきい値による実際の結果の精度算出をそれぞれの手法で独立に行った．Web を用いた語の連想関係抽出実験では，妥当なしきい値は 0.20 と決定し，その時 0.23 の F 値を得た．構築されたグラフの連結成分の大小による分類実験では，妥当なしきい値は 0.40 と決定し，その時偽陽性率が 0.00，偽陰性率が 0.20 の分類精度を得た．

Visualizing Typicality of Text Using Associative Relation between Words

Takeru Nakabayashi,[†1] Takayuki Yumoto,[†1]
Manabu Nii[†1] and Yutaka Takahashi

Recently, we collect information on the web. Then, we often choose text by typicality of the text. Therefore, we propose a method of visualizing typicality of text. This method finds atypical sentences and the reasons. We extract associative relation between words using web. Then, we make graphs based on the associative relation between words. Finally, we classify the graphs by that their shape, and we judge typicality of sentences. We evaluated our algorithm of extracting associative relation, and the average F-measure was 0.23. We also evaluated our algorithm of classifying graphs of associative relation, and the false positive rate was 0.00 and the false negative rate was 0.20.

1. はじめに

近年，Web が普及し誰でも容易に Web で情報収集することができるようになった．しかしながら，Web に存在する文章は様々な記述がされており，非典型的な文章も多い．そのため，情報リテラシーの低いユーザや，その事柄について知識が無いユーザは非典型的な文章を一般的な情報と勘違いをする．また，自ら文章の典型性を検証するにも，検索エンジンを用いて他の文章を取得し比較する等，多大な労力を費やす．たとえば Wikipedia の一文の “鉛筆とは筆記具，文房具の一種である．” という文は典型的であるが，Uncyclopedia の一文の “鉛筆とは簡易型ロケットの事である．” という文は非典型的である．我々は鉛筆についての知識があるため，これらの典型性を判断することができるが，知識が無い場合には判断が困難になる．このように，情報の取捨選択は文章が典型的か否かで判断することが多く，その行為は一般的に負担が大きい．そこで我々は，文章の典型的でない記述がされている箇所とその根拠または補助となる情報を視覚的に提示する手法を提案する．図 1 に構想しているインタフェースの例を示す．このようにして出力された文の典型性は，一般的でない意見や理論の発見にも利用でき，さらには信憑性判断の指標にも利用できると考えている．

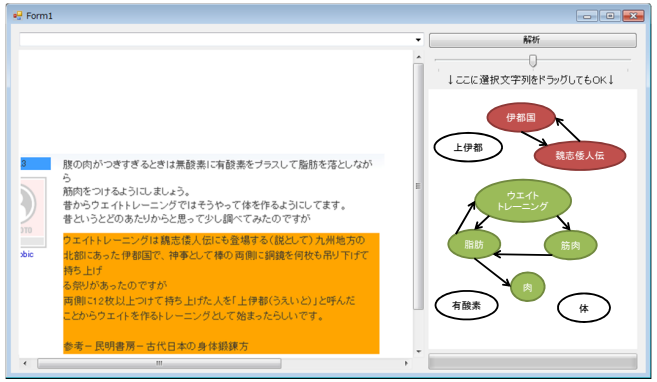

図 1 ユーザインタフェースイメージ図

†1 兵庫県立大学大学院工学研究科
Graduate School of Engineering, University of Hyogo

本研究では，文章を文単位で考え，文中に同時に使用されている語同士の関係が強ければ典型的な文であると考える．提案手法は，文中に出現する語同士の連想関係を Web を用いて推定する手法と，連想関係に基づいて構築されたグラフのパターン分類によって典型性を判断する 2 つの手法から構成する．語同士の連想関係では Web をコーパスとした確信度を用いて判定する．また，グラフパターンの分類については一文中での話題数に着目している．

2. 関 連 研 究

2.1 文書外の情報を用いた語の関係抽出

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[1]．このシステムは，検索結果の集約と Web ページの生成時間分布解析をすることで，高い精度で単文の信用度評価を可能としている．基本的なアプローチとして，彼らは語のフレーズを用いて Web から検索結果件数を取得し，信用度評価を行っている．例えば，“恐竜は 6500 万年前に絶滅した” というフレーズの場合，ユーザが “6500 万年” が正しいか知りたいとする．この場合，Web から “6500 万年”，“7000 万年”，“1 万年” といったような表現を抽出する．そして，抽出された表現と，その前後の表現を同時に用いて “恐竜 1 万年 絶滅” といったフレーズを作成した後，検索エンジンに問い合わせをする．その結果，Web での多数決が行われ，一般的に言われている典型的なフレーズが判明する．このように，フレーズ検索は精度の高い評価を得られる一方，表記ゆれに対応できないといった問題がある．我々は，Web における多数決という基本的な考えは同じであるが，“語” という単位で多数決を行い連想関係を抽出する．そうすることで，より汎用的な利用が可能であると考えている．

中山らは，Wikipedia のリンク解析を行うことでシソーラスの辞書を構築し，Wikipedia シソーラスを開発している[2)3)]．Wikipedia シソーラスはある語を入力とした時，関連の強い語が出力として得られるサービスである．Wikipedia シソーラスは Wikipedia 間のハイパーリンクの共起を用いて語間の関連度を算出し，関連の高い語を抽出する．その際，独自の pf-idf と呼ばれる値を用いる．これは，ハイパーリンクにおける TF-IDF の概念を取り入れた値であるため，非常に精度の高い結果が得られる．しかしながら，非常に関連の強い語のみ出力するため，抽出される関係が少なく，本研究で使用し難い．我々は，Web をコーパスとすることでより多くの関係が抽出可能となる語の関係抽出手法を目指す．

2.2 文書内の情報を用いた語の関係抽出

本研究では語の関係抽出に文書外の情報を用いる．そのため，本節で述べる関連研究は語の関係抽出と直接関係は無いが考え方やユーザへの提示方法を参考とした．

大澤らは著者独自の考えの主張を表すキーワード抽出法として，“KeyGraph” を開発している[4)]．KeyGraph は他の文書を用いることなく，文書単独でキーワード抽出を行う．KeyGraph は，一般語は語の出現頻度で，著者の主張を表す語は共起などの統計的指標を用いて関係を抽出し，抽出された関係によってエッジで結ばれグラフを構築する．構築されたグラフをユーザへ提示することで，著者の主張と他の語の関係を視覚的に把握することを可能としている．本研究では，語の関係を視覚的に得る方法として KeyGraph を参考とした．

また，松尾らは他のコーパスを用いない文書単独でのキーワード抽出法を提案している[5)]．彼らは，一文中における語の共起の統計情報を用いている．本研究では，一文中における語の共起という考え方を参考として，文書外の情報を利用した語の共起を手法に取り入れた．

彼らの手法について簡単に説明する．彼らは “文書全体で一文中における語の共起を調べた場合，語の共起に偏りはない” と仮説を立て検定を行い，仮説が棄却された共起の関係を抽出する．

本研究では，語の連想関係を抽出しそれを用いて有向グラフを構築することを目指す．これらの文書単独での語の関係抽出法は，注目している文書に大きく依存した語の関係が抽出される．そのため，本研究が目指す連想関係の抽出には向かないと考える．よって，本研究では文書外の情報を用いた語の関係抽出手法を提案する．

3. 語の連想関係に基づく文の典型性判定手法

本研究では，Web 上の文章を対象として，以下の 2 段階で典型性を解析する．また，図 2 に概要図を示す．

(1) Web をコーパスとして語の連想関係を推定しグラフを構築
(2) 語の連想関係に基づくグラフのパターン分類

なお，提案手法は文章を文単位で処理し，使用する語はサ変接続を除く名詞である．文から名詞を抽出する際には，形態素解析器である MeCab を用いる[⋆1]．

3.1 語の連想関係グラフモデル

本研究では，語の連想関係をグラフで表現し，それに基づき典型的な文と非典型的な文を分類する．本節では語の連想関係の推定とそのグラフ表現，構築されたグラフと文章典型性の関係について述べる．

⋆1 http://mecab.sourceforge.net/

Vol.2011-DBS-153 No.5
2011/11/3

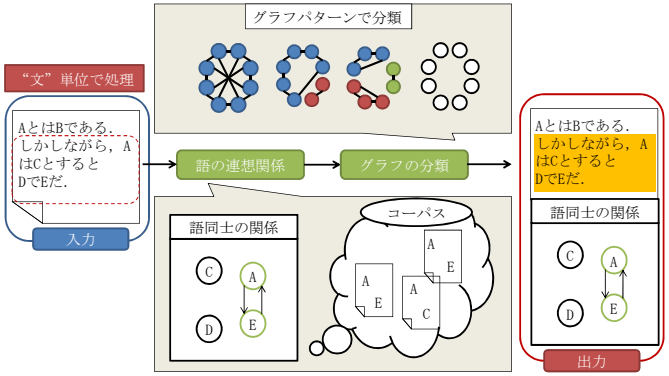

図 2 処理概要

3.1.1 語の連想関係とグラフ表現

ある語が与えられたとき，関係のある他の語が連想されることがある．たとえば，“ブラウザ” という語が与えられたとき，多くの人は別の語である “Internet Explorer” や “Mozilla Firefox” 等を連想する．人によってはそれらの語はお互いに連想関係にあると考えるが，我々は連想関係には方向性が存在すると考えている．例えば，“バナナ” から “黄色” は連想可能であるが，逆に “黄色” から “バナナ” は必ずしもそうではないと言える．これは，“バナナ” という語に対して “黄色” という語は非常に広く用いられているためである．つまり，集合の大きさに大きな差があり，包含関係が成り立つような場合にこのような片方向の連想関係があると考えられる．よって本研究では，(1) 式を満たす時，語 A から語 B が連想可能であるとする．

$$\theta < \frac{DF(A \wedge B)}{DF(A)} \tag{1}$$

ただし，θ は任意のしきい値，$DF(X)$ は語 X を含む文書数，$DF(X \wedge Y)$ は語 X と語 Y を同時に含む文書数である．提案手法では Web をコーパスとして用いる．具体的には Bing Search Engine API[⋆1]を利用して検索結果件数を取得し，これを計算に用いる．つまり $DF(X)$ は語 X の検索結果件数である．このように，ある語から他の語への方向性を持った連想関係が存在するとき，その連想関係を有向グラフで表現することができると考える．

⋆1 http://www.bing.com/toolbox/bingdeveloper/

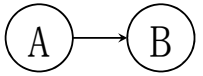

図 3 語 A から語 B が連想可能である場合のグラフ表現例

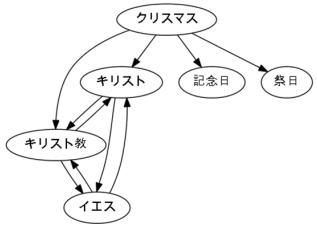

図 4 クリスマスについての記述から構築されたグラフ

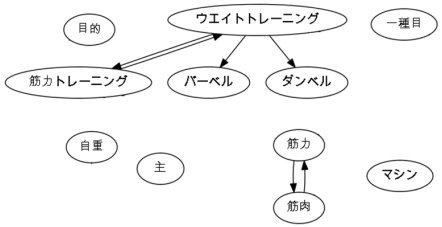

図 5 ウエイトトレーニングについての記述から構築されたグラフ

図 6 日本放送協会について虚偽が記述された文から構築されたグラフ

語 A から語 B が連想できるとき，$A \rightarrow B$ のように方向を持たせた表現ができる．図 3 に，連想関係のグラフ表現の例を示す．

3.1.2 文の典型性とグラフパターンの関係

予備調査として，前節で述べた語の連想関係の推定手法を用いて実際の文からグラフを構築した．以下に，調査結果の一部を図 4〜6 に示す．図を見ると，文によってグラフの構造パターンがいくつかあることがわかる．例えば図 4 では，すべてのノードが他のノードとエッジを張り “大きな一つ” の連結成分でグラフを構築している．文を構成している語同士が互いに連想できる関係であることから，“中心となる話題を一つ含む文” であると考えられる．しかしながら，図 5 では，各ノードは数本のみエッジを張り “小さな複数” の連結成分でグラフを構築している．これを図 4 の例と同様に考えると，“話題が複数存在し，中

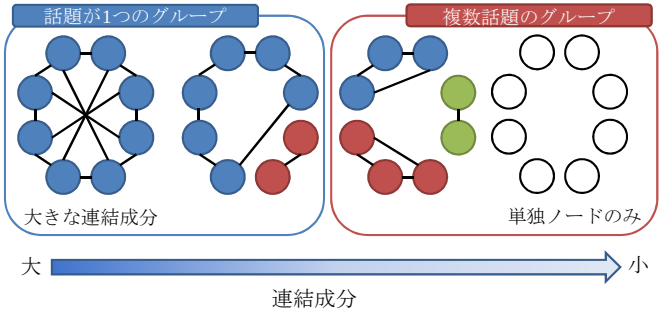

図 7 想定されるグラフ構造パターン

心の話題を決定することができない文” であると言える．また，図 6 では，全てのノードはエッジを張らず，“単独のノード” でグラフを構築している．これは極端な例で，本質的には図 5 の例と同様の文であると考えられる．

この予備調査より我々は，少なくとも “一文で話題が複数存在するような文は非典型的である” と考える．実際入力に用いた文の結果と比較すると，典型的な文は図 4 のようなグラフになり，非典型的な文は図 5，6 のようなグラフになる傾向があることがわかった．したがって，図 4 のようなグラフを構築する文は典型的な文とみなし，図 5 や図 6 のようなグラフを構築する文は非典型的な文に分類する．これらを一般化したグラフ構造のパターンを図 7 に示す．

図 7 に示したグラフ構造パターンは，図 4 のようなグループと，図 5，6 のようなグループに分類できる．本手法では分類の基準として，連結成分の大きさに着目する．具体的には，グラフ G の中でノード数が最大となる連結成分 C_G に着目した時，C_G がグラフ全体のどれくらいの割合を占めているかという基準で分類する．(2) 式を満たす時，“話題が一つのグループ” として分類する．

$$\eta \leq \frac{size(C_G)}{size(G)} \tag{2}$$

ここで，η はしきい値で $size(H)$ はグラフ H のノード数である．“話題が一つのグループ” と認められなかったグラフを “複数話題のグループ” として分類する．

3.2 グラフパターンに基づく典型性の可視化

本研究では，複数話題のグループに属するグラフを構築した文を非典型的な文として判断し，強調表示を行う．そうすることで，ユーザは非典型的な文を一目で認識でき，情報の取捨選択の負担が軽減できると考える．また，構築されたグラフはユーザが文章を読む際の補助的な情報として，同時に出力する．

図 1 で示した具体的な例で説明する．図の左側がユーザの主な閲覧領域となっており，右側がシステムによって構築されたグラフ表示領域である．右側に表示されている連想関係のグラフが根拠となって，閲覧領域のオレンジ色に強調されている部分が非典型的であるとユーザに提示している．この例では，表示されているグラフは “ウエイトトレーニング” についての連結成分と “魏志倭人伝” についての連結成分を保持している．“ウエイトトレーニング” についての連結成分が，このグラフの中で最大ノード数となる連結成分である．ここで，その連結成分のノード数は全体のノード数の 5 割程であるため，“複数話題のグループ” に分類される．以上の理由から，このグラフを構築した文をオレンジ色に強調している．

4. 評 価 実 験

提案手法について評価実験を行った．本研究で想定している入力は “文章” であるが，今回は文単位での処理を含め “語の連想関係の抽出” および “グラフパターンによる文の分類” の評価を行うため，文単位で評価を行った．まず，3.1.1 節で述べた手法で “語の連想関係の抽出” が理想通り行われるか評価実験を行った．また，3.1.2 節で述べた “グラフパターンによる文の分類” で，仮説通り分類が行われるか評価実験を行い，それぞれの手法が有効であるか検討した．実験には，以下に示す表 1 について記述された文を入力した．典型的な文として Wikipedia から，非典型的な文として Uncyclopedia からそれぞれ文を取得した．

実際に使用した文の一部を以下の表 2 および表 3 に示す．基本的にエントリの語を主語とし，端的にエントリについて説明している文を取得した．

4.1 語の連想関係抽出実験

3.1.1 節で述べた手法を用いて，表 2，3 の入力文から語の連想関係の抽出を行った．今回，用いる (1) 式のしきい値 θ を [0.1,1.0] の範囲で 0.1 刻みで変化させ，連想関係を抽出

表 1 入力テキストのエントリ一覧

アメリカ合衆国	マクドナルド	鉛筆
Mozilla Firefox	日本放送協会	クリスマス
Google	オブジェクト指向	Hyde
ウエイトトレーニング		

Vol.2011-DBS-153 No.5
2011/11/3

した．抽出された連想関係の結果の一部を表 4 に示す．

また，これらの連想関係が妥当であるか評価した．この時，$A \to B \to C$ のような連想関係の場合，$A \to B$ と $B \to C$ のように分割し連想関係のペアについて評価した．評価のために，実際に使用した語を用いて正解セットを作成した．正解セットの作成のため 6 人の被験者にアンケートをとり，本人にとって連想可能である語の対を記述させた．このうち過半数が挙げている連想関係を，正解の連想関係とした．正解の連想関係の一部を表 5 に示す．ただし，連想元が存在するが連想先のない語は，単独ノードを表している．評価指標には，連想関係の再現率，適合率および F 値を算出し用いた．再現率，適合率および F 値は主に検索エンジンの性能を示す指標として知られている．これらの指標 (特に F 値) で高い値を示せば，本手法で理想の連想関係を抽出できたといえる．

再現率 (Recall) はすべての正解の中で，システムが出力した正解の割合である．つまりシステムがどれだけ正解をカバーできているかを示す指標である．本手法では (3) 式のように定義する．

$$R = \frac{Num_c}{N_{answer}} \tag{3}$$

ここで，Num_c は正解セットと実験結果で一致した連想関係の数であり，N_{answer} は正解セットの連想関係の数である．

適合率 (Precision) はシステムが出力した中にどれだけ正解が含まれているか，つまりシステムの正解率である．よって本手法では (4) 式のように定義する．

表 2 入力文一覧 (Wikipedia)

記事エントリ	入力文
アメリカ合衆国	アメリカ合衆国は通称アメリカまた合衆国，米国とよばれ，北アメリカ大陸および北太平洋に位置する連邦共和国である．
鉛筆	鉛筆とは筆記具，文房具の一種であり，顔料を細長く固めた芯（鉛筆芯）を軸（鉛筆軸）ではさんで持ち易くしたもので，鉛筆の片側の末端部分を削って露出させた芯を紙に滑らせると，紙との摩擦で芯が細かい粒子になり紙に顔料の軌跡を残すことで筆記される．
クリスマス	クリスマスはイエス・キリストの降誕（誕生）を祝うキリスト教の記念日・祭日である．
ウエイトトレーニング	ウエイトトレーニングは，筋力トレーニングの一種目でバーベル，ダンベル，マシンまたは自重などを使い筋肉に負荷をかけ体を鍛えるトレーニングで主に筋力の増大，またはそれに伴う筋肉の増量などを目的とするトレーニングの総称である．

表 3 入力文一覧 (Uncyclopedia)

記事エントリ	入力文
アメリカ合衆国	アメリカ（飴梨花）は権利団体のひとつで，転じて彼らが著作権をもつ全ての物事を示す言葉となった．
鉛筆	鉛筆（えんぴつ）とは簡易型ロケットのことで，アメリカとロシアの宇宙戦争についての名言がある．
クリスマス	クリスマスとは国が進める少子化及び消費支出対策の一環である国民の日．
ウエイトトレーニング[⋆1]	ウエイトトレーニングは魏志倭人伝にも登場する伊都国で行われていた祭りで，両側に銅鏡を 12 枚以上つけて持ち上げた人を “上伊都” と呼んだことからウエイトを作るトレーニングとして始まったらしいです．

⋆1 http://oshiete.goo.ne.jp/qa/4518912.html

表 4 連想関係抽出結果 (左：Wikipedia，右：Uncyclopedia)：$\theta = 0.4$

クリスマス	
連想元	連想先
イエス	クリスマス
イエス	キリスト
イエス	キリスト教
キリスト	イエス
キリスト	キリスト教
キリスト教	イエス
キリスト教	キリスト
記念日	クリスマス
祭日	

クリスマス	
連想元	連想先
消費支出対策	クリスマス
消費支出対策	少子化
消費支出対策	一環
消費支出対策	国民
一環	国民
少子化	国民

表 5 連想関係正解セット (左：Wikipedia，右：Uncyclopedia)

クリスマス	
連想元	連想先
クリスマス	キリスト
クリスマス	キリスト教
クリスマス	記念日
クリスマス	祭日
イエス	キリスト
イエス	キリスト教
キリスト	イエス
キリスト	キリスト教
キリスト教	キリスト
キリスト教	イエス

クリスマス	
連想元	連想先
クリスマス	
少子化	
消費支出対策	
一環	
国民	

$$P = \frac{Num_c}{N_{result}} \tag{4}$$

ここで，N_{result} は手法により抽出された連想関係の数である．

F 値は再現率と適合率の調和平均であり，(5) 式のように定義される．

$$F = \frac{2 \times R \times P}{R + P} = \frac{2 \times Num_c}{N_{answer} + N_{result}} \tag{5}$$

ここで，提案手法が “他の語の共起尺度を用いて抽出される連想関係” と比べ，精度がよいか調査実験を行った．今回は，Jaccard 係数，Simpson 係数，Lift を用いて提案手法と同様に，語の連想関係の抽出を行い，F 値を算出した．Jaccard 係数，Simpson 係数は Web を用いた語の共起尺度として知られている[6]．Lift は相関ルールの抽出で用いられ，事象 A と事象 B の関連性を知ることができる[7]．それぞれの定義式を (6)〜(8) 式に示す．

$$Jaccard(A, B) = \frac{DF(A \wedge B)}{DF(A \vee B)} \tag{6}$$

ここで，$DF(A \wedge B)$ は，語 A と語 B を同時に含む文書数，$DF(A \vee B)$ は少なくとも語 A または語 B どちらかを含む文書数である．

$$Simpson(A, B) = \frac{DF(A \wedge B)}{min(DF(A), DF(B))} \tag{7}$$

ここで，$min(DF(A), DF(B))$ は，$DF(A)$，$DF(B)$ の文書数の内，少ない方の文書数である．

$$lift(A \Rightarrow B) = \frac{DF(A \wedge B) \times |D|}{DF(A) \times DF(B)} \tag{8}$$

ここで，$A \Rightarrow B$ は “語 A が出現する場合，語 B も同文書内で出現する” というルールで，$|D|$ は文書集合全体の数である．

この時，しきい値 θ は [0.1,1.0] の区間で 0.1 ずつ変化させ連想関係の抽出を行い，それぞれのしきい値で F 値を算出した．ただし，Lift で用いる $|D|$ は 500,000,000 に設定し，算出された Lift の値は対数をとった．結果を表 6 と図 8，9 に示す．

表 6 から，F 値の最大値に大きな差は見られないが，若干 Lift を用いた手法の精度が良

表 6 ベースラインとの比較

	Confidence	Jaccard	Simpson	Lift
F 値の最大値	0.23	0.13	0.27	0.30
しきい値 θ	0.20	0.10	0.10	1.80

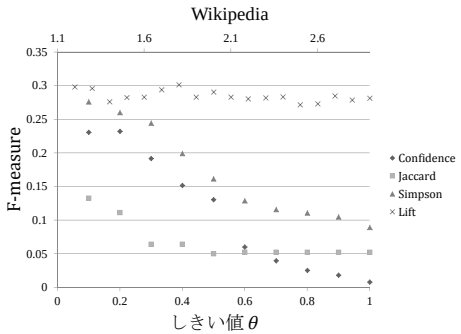

図 8 他の共起尺度との比較 (Wikipedia)

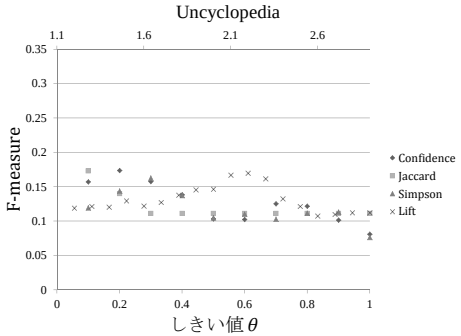

図 9 他の共起尺度との比較 (Uncyclopedia)

いことがわかった．しかしながら，Lift の最大値は定まっておらず，しきい値の θ 決定が難しい．今回は，Lift においてのみしきい値 θ の値域を [1.2,3.0] とし，しきい値 θ を変化させ実験を行った．Lift を用いた手法は，図 8 を見る限りではしきい値を変化させても F 値平均の値も大きな変化がないことがわかる．図 9 では少し変化しているが，その値は小さい．また，図 8 から，全てのしきい値において提案手法よりも Simpson 係数を用いた連想関係抽出手法および Lift を用いた連想関係抽出手法のほうが精度が良いことがわかった．しかしながら，本研究で考えている，語の連想関係の方向性はこれらの尺度を用いると表現できない．よって今後，提案手法とこれらを絡めた手法を考案することで連想関係の方向性を維持しつつ精度の向上を目指す．ここで，提案手法である確信度の F 値に注目すると，本手法に用いるしきい値 θ は 0.2 が望ましいと言える．

また，提案手法における各しきい値ごとの再現率と適合率の平均を表 7 に示す． 再現率と共に適合率も増加し，一定値を超えると適合率が減少するといった傾向が見られる．これは本手法で抽出される連想関係の数が，しきい値 θ の増加に対して単純増加 (または単純減少) するわけではないことが原因であると考えられる．また，結果から再現率，適合率共に値が低いといった問題も見られる．ここで，$\theta = 0.4$ の時の各文の評価値を表 8，9 示す．

表 8 を確認すると F 値において 0.63 と高い値を示している文も存在した．ここで比較的高い値を示した “クリスマス (Wikipedia)” の実験結果と正解セットに着目すると，それぞれ連想関係の数が 10 前後と差が小さく，単独のノードがほとんどないことがわかった．

一方で，再現率，適合率および F 値全てにおいて 0 となった “クリスマス (Uncyclopedia)” の実験結果と正解セットを比較すると，明らかに連想関係の差異がみられた．加えて，正解

セットでは単独のノードが多いが実験結果ではほぼ単独のノードは存在しないといった差異があった．これは，例えば “鉛筆軸” と “軸” のような語の一部に片方の語が含まれているような複合語は検索結果として出現し，連想関係として抽出されるが，人間には連想され難く正解として扱われなかったことが原因だと考えられる．また，Web をコーパスとして用いている以上，避けようのないノイズが結果に大きく影響していると考えられる．

表 7 各しきい値における連想関係の再現率・適合率平均 (左：Wikipedia，右：Uncyclopedia)

しきい値	再現率	適合率
0.10	0.273	0.200
0.20	0.240	0.225
0.30	0.180	0.204
0.40	0.130	0.181
0.50	0.110	0.160
0.60	0.052	0.070
0.70	0.032	0.050
0.80	0.020	0.034
0.90	0.013	0.026
1.00	0.006	0.010

しきい値	再現率	適合率
0.10	0.212	0.125
0.20	0.187	0.162
0.30	0.162	0.153
0.40	0.138	0.138
0.50	0.102	0.105
0.60	0.102	0.103
0.70	0.127	0.123
0.80	0.127	0.117
0.90	0.107	0.096
1.00	0.086	0.075

表 8 語の連想関係抽出の評価 (Wikipedia)

	アメリカ合衆国	マクドナルド	鉛筆	Mozilla Firefox	日本放送協会
再現率	0.43	0.17	0.24	0.00	0.062
適合率	0.50	0.25	0.17	0.00	0.025
F 値	0.46	0.20	0.20	0.00	0.035

	クリスマス	Google	オブジェクト指向	Hyde	ウエイトトレーニング	平均
再現率	0.60	0.17	0.40	0.00	0.47	0.25
適合率	0.67	0.14	0.22	0.00	0.18	0.22
F 値	0.63	0.15	0.29	0.00	0.25	0.22

表 9 語の連想関係抽出の評価 (Uncyclopedia)

	アメリカ合衆国	マクドナルド	鉛筆	Mozilla Firefox	日本放送協会
再現率	0.25	0.50	0.43	0.22	0.00
適合率	0.080	0.29	0.30	0.07	0.00
F 値	0.12	0.36	0.35	0.11	0.00

	クリスマス	Google	オブジェクト指向	Hyde	ウエイトトレーニング	平均
再現率	0.00	0.00	0.33	0.40	0.25	0.24
適合率	0.00	0.00	0.23	0.29	0.22	0.15
F 値	0.00	0.00	0.27	0.33	0.24	0.18

これらが要因となり最適なしきい値 $\theta = 0.2$ においても平均精度 (F 値) が低くなったと言える．この評価値から，語の連想関係抽出手法の改善の必要がある．現段階では，正解セットに含まれない実験結果の連想関係で，概ね正しいと感じられる関係もいくつか確認できることから，正解セット作成方法に問題があったと考えられる．したがって，正解セットの見直しを行い再検証する必要がある．

4.2 グラフパターンによる文の分類

3.1.2 節で述べた手法を用い，作成した正解セットの連想関係グラフをグラフパターンにより分類した．今回，しきい値 η を [0.1,1.0] の範囲で 0.05 刻みで変化させ分類を行った．表 10，11 に (2) 式の分類基準値と $\eta = 0.7$ とした時の分類結果を示す．

また，得られた分類結果と仮説を比較し評価した．仮説では，“話題が一つのグループ” に分類されるグラフは典型的な文，“複数話題のグループ” に分類されるグラフは非典型的な文として考えた．

評価のため，偽陽性率 (False positive rate) と偽陰性率 (False negative rate) を算出する．偽陽性率と偽陰性率は統計上の過誤を表す用語で，スパムメール判定などのフィルタリングシステムや検査などでよく用いられる[8]．今，典型的な文を中心として考えると，典型的な文を非典型的な文と誤って分類することを “偽陰性”，非典型的な文を典型的な文と誤って分類することを “偽陽性” と呼ぶ．偽陽性率 α，偽陰性率 β はそれぞれ (9) 式，(10)

表 10 分類基準値と分類結果 (Wikipedia)：$\eta = 0.7$

	アメリカ合衆国	クリスマス	Google	日本放送協会	ウエイトトレーニング
分類基準値	0.71	1.00	0.67	0.42	0.55
分類	典型	典型	非典型	非典型	非典型

	マクドナルド	鉛筆	オブジェクト指向	hyde	Mozilla Firefox
分類基準値	0.80	0.43	0.67	0.75	1.00
分類	典型	非典型	非典型	典型	典型

表 11 分類基準値と分類結果 (Uncyclopedia)：$\eta = 0.7$

	アメリカ合衆国	クリスマス	Google	日本放送協会	ウエイトトレーニング
分類基準値	0.25	0.00	0.17	0.00	0.29
分類	非典型	非典型	非典型	非典型	非典型

	マクドナルド	鉛筆	オブジェクト指向	hyde	Mozilla Firefox
分類基準値	0.60	0.29	0.44	0.50	0.30
分類	非典型	非典型	非典型	非典型	非典型

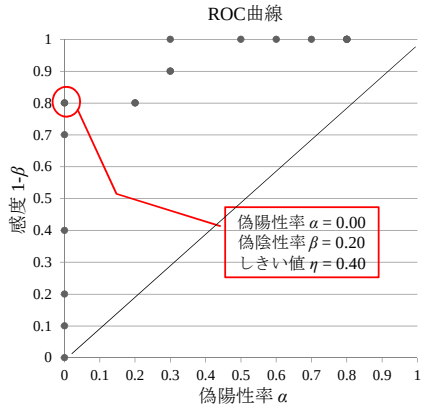

図 10 ROC 曲線

式のように定義される．

$$\alpha = \frac{number\ of\ false\ positives}{number\ of\ negative\ instance} \tag{9}$$

$$\beta = \frac{number\ of\ false\ negative}{number\ of\ positive\ instance} \tag{10}$$

図 10 に，しきい値 η 毎に算出した偽陽性率および偽陰性率を用いて描いた ROC 曲線を示す． ROC 曲線は分類器の性能を評価する際に用いられる．縦軸を $1-\beta$，横軸を α とした時，分類器の出力した結果が $y = x$ の直線より左上に寄れば寄るほど性能が良いことを示す．図 10 の曲線では，かなり左上に寄っていることから非常に良い分類結果が得られたことがわかる．また，最も誤らず分類を行ったしきい値 η は 0.4 であった．よって本手法ではしきい値 η は 0.4 が妥当であると言える．

5. おわりに

本稿では，情報の取捨選択支援のために，語の連想関係を用いた典型性の可視化手法を提案した．本手法は大きく分けて 2 段階の処理で典型性の可視化を行う．まず 1 段階目では，Web をコーパスとして語の連想関係を抽出し，その関係を用いて有向グラフを構築する．この際，Web を用いた確信度を使う．2 段階目では，構築されたグラフの連結成分の大小によるグループ分類を行う．分けられたグループにより，典型的な文か否かを判定し，非典型的と判定された文を強調する．強調された文はその根拠として構築したグラフと同時に出力する．これら強調とグラフを見ることでユーザは一目で文の典型性を把握できると考える．実験を行った結果，語の連想関係の抽出手法に精度の問題が見られた．しかしながら，グラフの連結成分の大小による文の分類手法は非常に高い精度を得られたため，連想関係の抽出手法の改善を行うことで，最終出力において十分な精度が期待できる．具体的には，Lift を考慮した確信度による語の連想関係抽出を行うことで精度の改善を図る．

参 考 文 献

1) 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, Vol.6, No.1, pp.53–56 (2007).
2) 中山浩太郎，原　隆浩，西尾章治郎：Wikipedia マイニングによるシソーラス辞書の構築手法，情報処理学会論文誌，Vol.47, No.10, pp.2917–2928 (2006).
3) 伊藤雅弘，中山浩太郎，原　隆浩，西尾章治郎：Wikipedia のリンク共起性解析によるシソーラス辞書構築，情報処理学会論文誌，Vol.48, No.SIG19(TOD 36), pp.39–49 (2007).
4) 大澤幸生，ネルス E. ベンソン，谷内田正彦：KeyGraph：語の共起グラフの分割・統合によるキーワード抽出，電子情報通信学会論文誌，Vol.J82-D-1, No.2, pp.381–400 (1999).
5) 松尾　豊，石塚　満：語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム，人工知能学会論文誌，Vol.17, pp.217–223 (2002).
6) Frakes, W.B. and Baeza-Yates, R.A.: *Information Retrieval: Data Structures & Algorithms*, Prentice-Hall (1992).
7) 元田　浩，津本周作，山口高平，沼尾正行：データマイニングの基礎，オーム社 (2006).
8) Manning, C.D., Raghavan, P. and Schuetze, H.: *Introduction to Information Retrieval*, Cambridge University Press, anniversary. edition (2008).