Vol.2018-DBS-167 No.19
Vol.2018-IFAT-132 No.19
2018/9/13

# 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

小山 雄也[1]　湯本 高行[1]　礒川 悌次郎[1]　上浦 尚武[1]

**概要**：Web 検索では，検索クエリを入力するとクエリに関する情報を手軽に手に入れることができるが，その一方で，検索結果に信頼性の低い情報が表示されることがある．特に利用者が検索クエリに関して知識がない場合，信頼性の判断基準がほとんどないという問題が存在している．そこで本研究では，信頼性の高い情報はクエリと関連の深い語で構成されていると考え，クエリと名詞の関連性に注目する．クエリと関連の深い名詞はクエリの出現する文章に共に出現する頻度が高いと考え，名詞に対するクエリの共起確率を用いたクエリと名詞の関連性の推定を目的とする．その際，共起確率に対する名詞の文書数の影響を低減するため，予測値に対する実測値の比を関連性の推定に用いる．また，関連性を定量化した関連度の被験者評価値をクラウドソーシングを用いて収集し，実際の関連度と提案手法で算出した値との比較を行う．

## 1. はじめに

近年，スマートフォンやタブレット型端末の普及により Web 検索の機会が増加し，場所や時間を選ばずに誰でも利用できるようになっている．Web 検索では利用者が検索クエリを入力すると，クエリに関連する Web サイトのランキングが表示され，そのランキングから見たいと思う Web サイトを利用者が自ら選択することで，情報が表示される．例えば，利用者がクエリとして「認知症」を入力し検索を行うと，認知症に関する Web ページのランキングが表示され，そのランキングから Web ページを選択することで認知症の原因や症状，予防法などを知ることができる．このように，Web 検索により，医療の知識がない人でも簡単に病気に関する情報を得ることが可能である．一方で，Web 上には不正確であったり根拠がないなどの信頼性の低い情報も多く存在している．例えば，医療関係の Web サイトであっても，医師や薬剤師などの専門家の監修を受けていない場合が存在している．しかし，利用者が検索クエリに関して知識がない場合，信頼性の基準がほとんどないため，その情報が信頼できるかどうかを判断することは難しい．

この課題へのアプローチとして，本研究ではクエリと名詞の関連性に注目する．信頼性の高い情報はクエリと関連の深い名詞で構成されていると考え，クエリと名詞の関連性の推定を行う．その際，クエリと関連の深い名詞はクエリの出現する文章に共に出現する可能性が高いと考え，名詞に対するクエリの共起確率を用いて関連性を推定する．なお，共起確率に対する名詞の文書数の影響を低減するため，予測値に対する実測値の比を関連性の推定に用いる．

[1] 兵庫県立大学大学院工学研究科

## 2. 関連研究

Web 検索の信頼性判断支援に関する先行研究として，山本らの研究 [1][2] や Akamine らの研究 [3] などがある．文献 [1] では曖昧な知識の真偽を Web 検索を用いて調べるタスクの被験者実験を行い，その際に重要視する項目を調べることで，ページの評価および，知識の信頼性支援の方法を提案している．また，文献 [2] ではクエリに関する情報に対する反証の文を提示することで，信憑性指向の Web 検索を支援することを提案している．文献 [3] では，Web ページの内容，発信元，表面的特徴に関する分析結果をまとめて提示することで信頼性の判断支援を行っている．本研究では，入力されたクエリの Web ページ中の名詞との関連性を信頼性の指標として用いる点でこれらの先行研究とは異なっている．

また，語と語の関係を算出する研究には検索エンジンを用いる Cilibrasi らの研究 [4] や，概念ベースを用いる渡辺らの研究 [5][6] などがある．文献 [4] では，Google の検索エンジンの検索結果数を用いて語と語の意味の関係性を算出しており，算出式として正規化 Google 距離を提案している．文献 [5][6] では，国語辞書の見出し語を概念と見なし，語義文に含まれる自立語を属性と考えた，概念と属性集合で構成される概念ベースを関連度の算出に用いている．概念ベースの属性には出現頻度を基に重みが付与されており，文献 [5] では属性集合の一致度を関連度とし，文

© 2018 Information Processing Society of Japan

献 [6] では文献 [5] の一致度に加えて概念語の概念ベースでの共出現を考慮して関連度を算出している．本研究では名詞と文書の関係を保存したデータベースにおける，語と語の共起確率から関連度を算出している点でこれらの研究とは異なっている．

## 3. 関連性推定手法

本研究ではクエリと関連の深い名詞はクエリの出現する文章に共に出現する頻度が高いと考え，名詞に対するクエリの共起確率を用いて関連性を定量化した関連度を算出する．共起確率の算出には後述する名詞-文書データベースを用いる．また，算出した関連度を被験者評価値と比較することで評価を行う．

### 3.1 名詞に対するクエリの共起確率

ある語 $w$ の出現確率 $P(w)$ を，語を含む文書数 $|D_w|$ と全文書数 $|D|$ の商と定義し，クエリと名詞の関係を名詞 $n$ に対するクエリ $q$ の共起確率 $P(q|n)$ を式 (1) で定義する．

$$P(q|n) = \frac{|D_{q\cap n}|}{|D_n|} \tag{1}$$

式 (1) より，$P(q|n)$ は名詞が単体で出現する確率と比較してクエリと名詞が同じ文書に出現する確率が高いほど値が高くなる．なお，式 (1) において，$|D_n|$ が $|D_{q\cap n}|$ と比較してきわめて大きいため，式 (2) に示す対数共起確率 $log_{10}P(q|n)$ を関連度として用いる．

$$log_{10}P(q|n) = log_{10}\frac{|D_{q\cap n}|}{|D_n|} \tag{2}$$

ここで，式 (2) を変形したものを式 (3) に示す．

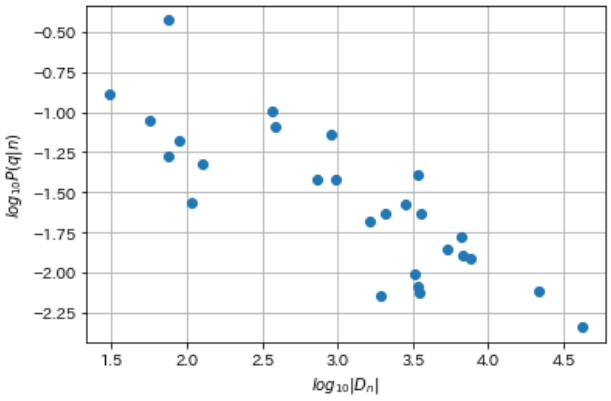
$$log_{10}P(q|n) = log_{10}|D_{q\cap n}| - log_{10}|D_n| \tag{3}$$

式 (3) より，右辺の $log_{10}|D_n|$ がきわめて大きいと，$log_{10}P(q|n)$ が $log_{10}|D_n|$ のみによって決まる可能性がある．そこで，事前調査として，$log_{10}P(q|n)$ に対する $log_{10}|D_n|$ の影響を調査した．実際のクエリ「認知症」とクエリに関する文に出現する 27 語の名詞の $log_{10}|D_n|$ と $log_{10}P(q|n)$ の関係を散布図で図 1 に示す．図 1 より，$log_{10}|D_n|$ が上昇すると，$log_{10}P(q|n)$ が減少する傾向が確認できる．この場合，あるクエリと低頻度の語及び高頻度の語との関係を比較したとき，常に低頻度の語の方が関連度が高くなるという問題が存在する．そこで，本研究では $log_{10}|D_n|$ から $log_{10}P(q|n)$ を予測し，予測値を用いて $log_{10}P(q|n)$ に対する $log_{10}|D_n|$ の影響の補正を行う．

### 3.2 予測比を用いた関連度

#### 3.2.1 区間平均値を用いた共起確率の予測

共起確率の予測には，まず，あるクエリ $q$ と共起するすべての名詞 $n$ の対数共起確率 $log_{10}P(q|n)$ を名詞-文書

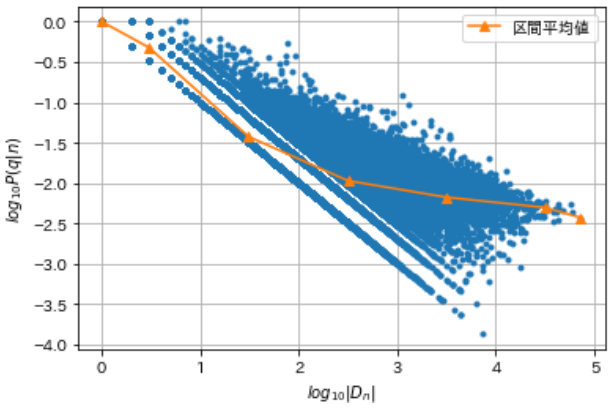



クエリ：認知症

**図 1** $log_{10}|D_n|$ と $log_{10}P(q|n)$ の関係

データベースより算出する．その後，名詞の文書数の対数 $log_{10}|D_n|$ を $[0,1),[1,2),[3,4),[4,5)$ の 5 つに分割し，分割した各区間の $log_{10}P(q|n)$ 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，平均値をつないで折れ線グラフを作成する．例えば，$log_{10}|D_n|$ が $[3,4)$ の区間の平均値は $log_{10}|D_n| = 3.5$ における値とする．ある $log_{10}|D_n|$ における折れ線の値を $log_{10}|D_n|$ での予測の値 $log_{10}\hat{P}(q|n)$ とする．

実際にクエリ「認知症」において，クエリと共起する全ての名詞の $log_{10}|D_n|$ と $log_{10}P(q|n)$ の関係及び，区間平均値をつないだ折れ線グラフを図 2 に示す．なお，折れ線の両端にはクエリと共起する全ての名詞のうち，それぞれ $log_{10}|D_n|$ が最小，最大の名詞の $log_{10}P(q|n)$ の値を用いる．




クエリ：認知症

**図 2** 区間平均値をつないだ折れ線グラフ

$|D_n|$ が小さい低頻度語では，共起数 1 回の違いで共起確率が大きく異なるため正確な推定が困難である．そのため，本研究では $|D_n| \leq 10$ の語では関連度推定を行わない．また，対数を用いるため $|D_{q\cap n}| = 0$ の語も同様に関連度推定を行わない．

#### 3.2.2 予測値に対する実際の共起確率の比

実際の共起確率 $P(q|n)$ の予測値 $\hat{P}(q|n)$ に対する比の対

© 2018 Information Processing Society of Japan

表 1　評価実験に使用した入力クエリ

| | | | |
|---|---|---|---|
| 北アメリカ | マダガスカル | 北極 | 冷蔵庫 |
| 自動車 | コンピュータ | エアバッグ | 重力 |
| 進化論 | 大気圏 | 原子力 | 認知症 |
| ダイエット | 睡眠 | プラシーボ | 不眠症 |
| 酵素 | ラグビー | 駅弁 | ユニクロ |

数を $log_{10}PR(q|n)$ と定義すると，式 (4) で表せる．

$$log_{10}PR(q|n) = log_{10}\frac{P(q|n)}{\hat{P}(q|n)} \tag{4}$$

また，式 (4) は式 (5) のように変形できる．

$$log_{10}PR(q|n) = log_{10}P(q|n) - log_{10}\hat{P}(q|n) \tag{5}$$

式 (5) より，$log_{10}PR(q|n)$ は予測の値からの実際の値の距離をあらわしており，予測値に対して実際の値が高いほど値が高くなる．$log_{10}\hat{P}(q|n)$ は $log_{10}|D_n|$ によって変化するため，$log_{10}P(q|n)$ の $log_{10}|D_n|$ の影響を補正することが可能であると考えられる．そこで，本研究では $log_{10}PR(q|n)$ を関連度として用いる．

## 4. 評価実験

本研究ではクエリと名詞の関連度を共起確率を用いて算出した．評価実験では入力に表 1 のクエリを用いた．クエリには，身近な事柄である単語を 20 語用いている．また，20 語のクエリに対して 5 文ずつの 100 文を Wikipedia から抽出し，各文に出現する 681 語との関連度を算出した．なお，本研究では低頻度語の関連度算出は行わないため $|D_n| > 10$ であり，$|D_{q\cap n}| > 0$ の 623 語を使用した．

### 4.1 データセット

#### 4.1.1 名詞-文書データベース

本研究では，クエリと名詞の共起確率を名詞と文書関係を保存したデータから算出する必要がある．そこで，はてなブックマークから名詞-文書データベースを構築する．URL から本文を抽出し，抽出した Web ページの本文に対して MeCab[*1][7] で形態素解析を行って名詞を抽出する．抽出した名詞と URL を ID で紐付けして名詞の出現数をカウントすることで，名詞-文書データベースを構築する．なお，固有名詞に対応するため，MeCab で使用する辞書には mecab-ipadic-neologd[*2]を用いる．

2014 年 4 月 24 日から 2017 年 1 月 27 日までの期間のはてなブックマークのホットエントリーからランダムに記事を抽出し名詞-文書データベースを構築した．構築したデータベースの規模を表 2 に示す．

#### 4.1.2 関連度の被験者評価値

提案手法の関連度を評価するために，Yahoo!クラウドソーシングを用いて，クエリとクエリに関する文に出現する名詞との関連度に関する被験者評価値データを収集した．実際に用いた質問の例を図 3 に示す．関連度の評価値として，名詞の組の関係を以下に示す質問で評価してもらい，a 以外の 4 つの選択肢に関係の深さの昇順に 1〜4 の評価値をつけ，選択された評価値の平均を用いている．その際，1 つの語の組に対して 10 人に調査を行った．

表 2　構築した名詞-文書データベースの規模

| | |
|---|---|
| URL 数 | 160,602 |
| 名詞数 | 1,537,808 |
| 名詞-URL の対応関係数 | 25,788,306 |

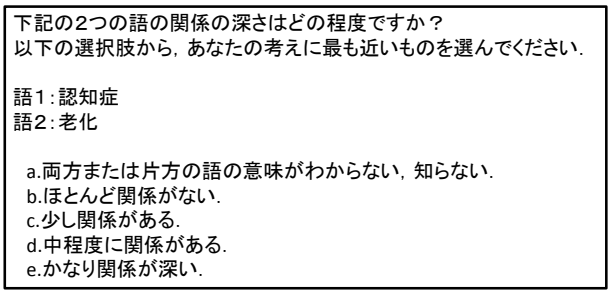



図 3　クラウドソーシングを用いた語の関連性に関する質問例

### 4.2 対数共起確率と予測比を用いた関連度の比較実験

#### 4.2.1 評価方法

本研究の提案手法である，3.2 節で示した予測比を用いた関連度に対して，3.1 節で示した対数共起確率をベースラインとして比較を行う．被験者評価値との相関が高い方が関連性の推定に適していると考え，対数共起確率及び予測比を用いた関連度と被験者評価値との相関係数をクエリごとに比較する．

#### 4.2.2 結果と考察

被験者評価値と対数共起確率及び，予測比を用いた関連度の相関係数を棒グラフで図 4 に示す．なお，予測比を用いた関連度の相関係数が対数共起確率の相関係数を下回っているクエリを灰色で表示している．図 4 より，予測比を用いた関連度は対数共起確率と比較して，15 クエリで相関係数が高くなっている．

相関係数が上昇したクエリ「北アメリカ」に関して，$log_{10}|D_n|$ ごとに色分けした対数共起確率と被験者評価値との関係を図 5 に，予測比を用いた関連度と被験者評価値の関係を図 6 にそれぞれ示す．図 5 より，ベースラインでは $log_{10}|D_n|$ の値が低い丸，バツのデータは $log_{10}P(q|n)$ の値が高い範囲に集中しており，反対に，$log_{10}|D_n|$ の値が高い三角と四角のデータは $log_{10}P(q|n)$ の値が低い範囲に集中している．このことから，ベースラインでは $log_{10}|D_n|$ の区間によって $log_{10}P(q|n)$ の値の高さがほとんど決定していることが確認できる．一方，図 6 より，提案手法では，

[*1] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://taku910.github.io/mecab/
[*2] 形態素解析辞書 mecab-ipadic-neologd: https://github.com/neologd/mecab-ipadic-neologd/wiki/Home.ja

© 2018 Information Processing Society of Japan

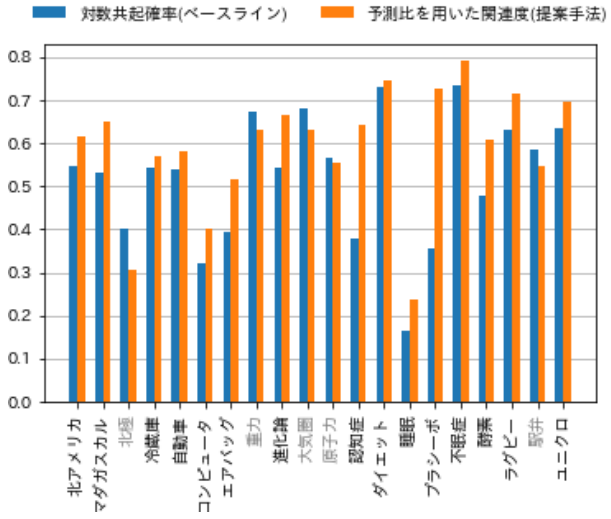
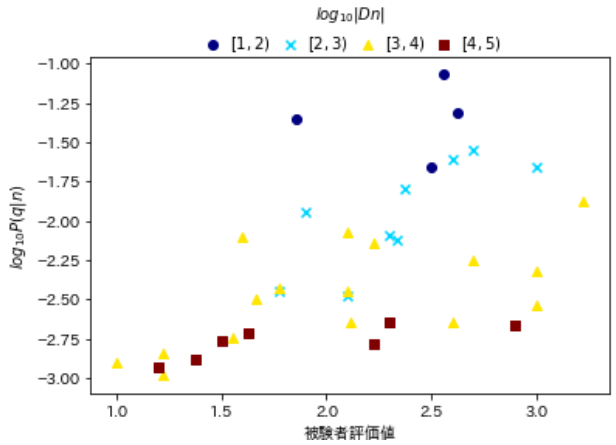



図 4　クエリごとの関連度と各共起確率との相関係数

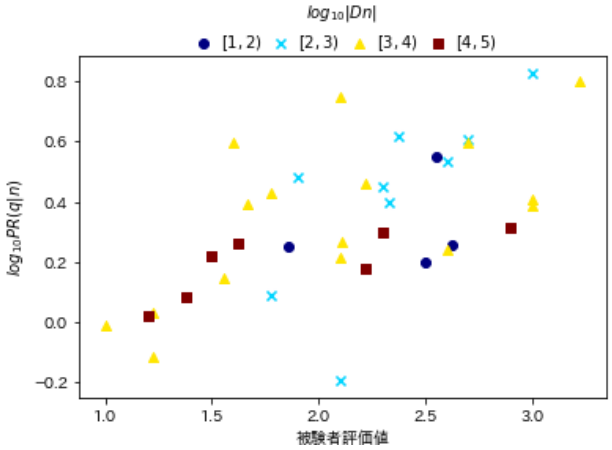
$log_{10}PR(q|n)$ が $log_{10}|D_n|$ の区間によらず広い範囲に分布している．このことから，予測比を用いた関連度は対数共起確率における $log_{10}|D_n|$ の影響を補正できており，これによって相関係数が上昇したと考えられる．




クエリ：北アメリカ

図 5　$log_{10}|D_n|$，対数共起確率，被験者評価値の関係

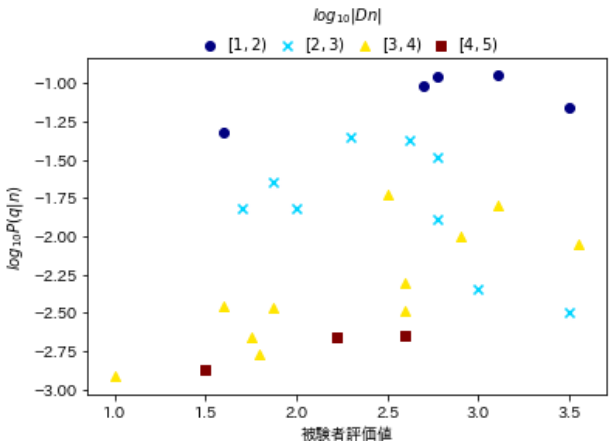
一方，相関係数が減少したクエリ「北極」に関して，$log_{10}|D_n|$ ごとに色分けした対数共起確率と被験者評価値の関係を図 7 に，予測比を用いた関連度と被験者評価値の関係を図 8 にそれぞれ示す．図 7, 8 より，クエリ「北極」でも「北アメリカ」と同様にベースラインでは $log_{10}|D_n|$ の区間によって $log_{10}P(q|n)$ の値の高さがほとんど決定しており，提案手法では $log_{10}|D_n|$ によらず広く分布していることが確認できる．また，図 8 より，予測比を用いても被験者評価値と関連度の傾向から離れている名詞が存在することが確認できる．例えば，被験者評価値が 3.5 付近で $log_{10}PR(q|n)$ の値がかなり低い名詞が存在している．そこで，傾向から離れている名詞を調査するため，予測比を用いた関連度と被験者評価値の関係に名詞を付与した散布図を図 9 に示す．図 9 より，「極地」や「北」など「北極」と




クエリ：北アメリカ

図 6　$log_{10}|D_n|$，予測比を用いた関連度，被験者評価値の関係

関係が深いと思われる名詞の $log_{10}PR(q|n)$ が高くなり，一般的な名詞である「そのもの」や「出現」で $log_{10}PR(q|n)$ が低くなっている．一方，「磁石」や「方位」が被験者評価値に対してかなり低くなっていることが確認できる．この原因として，これらの名詞が「北極」の話題だけで用いられる訳ではない一般的な名詞であることが挙げられる．例えば，「北極」の話題で「磁石」が出現する頻度は「磁石」の話題で「北極」が出現する頻度と異なると考えられる．そのため，名詞に対するクエリの共起確率で方向を考えた共起確率を用いている本手法の値とクエリと名詞を区別していない被験者評価値とで差が生じたと考えられる．今後の課題として，双方向の共起確率を考慮した関連度の開発が挙げられる．

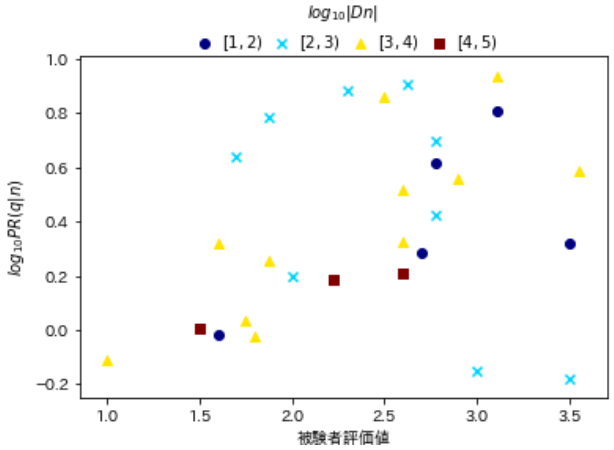



クエリ：北極

図 7　$log_{10}|D_n|$，対数共起確率，被験者評価値の関係

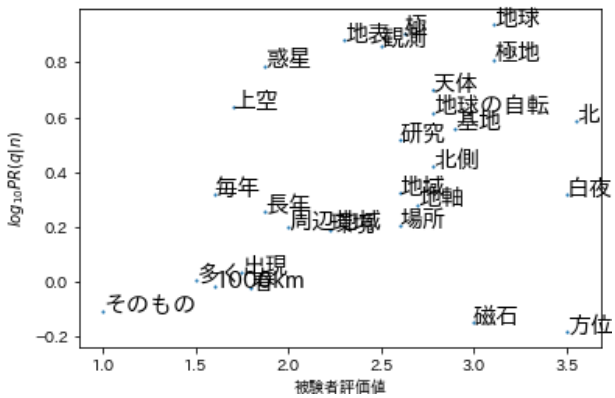
どちらの手法でも相関係数の低い「睡眠」に関して，$log_{10}|D_n|$ ごとに色分けした対数共起確率と被験者評価値の関係を図 10 に，予測比を用いた関連度と被験者評価値の関係を図 11 にそれぞれ示す．図 10,11 より，クエリ「睡眠」ではベースラインと提案手法の分布の違いがほとんど見られないことが確認できる．また，予測比を用いた関連

© 2018 Information Processing Society of Japan



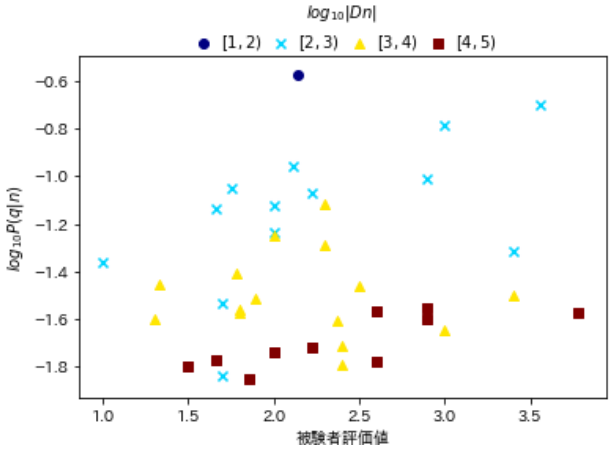
クエリ：北極

図 8　$log_{10}|D_n|$，予測比を用いた関連度，被験者評価値の関係



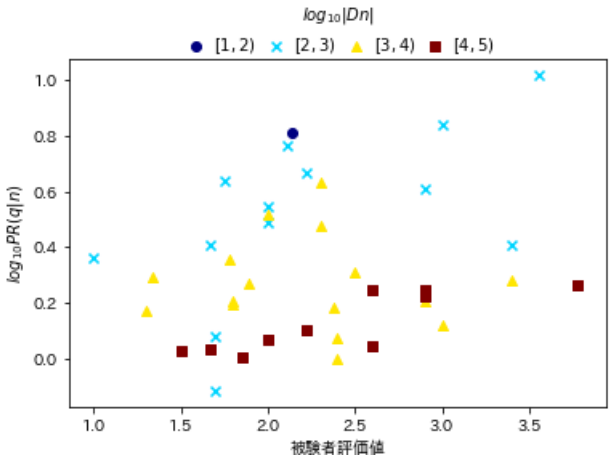
クエリ：北極

図 9　名詞を付与した予測比を用いた関連度と被験者評価値の関係



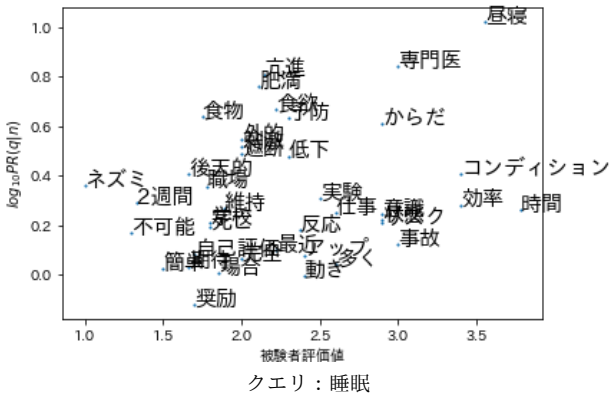
クエリ：睡眠

図 10　$log_{10}|D_n|$，対数共起確率，被験者評価値の関係



クエリ：睡眠

図 11　$log_{10}|D_n|$，予測比を用いた関連度，被験者評価値の関係



クエリ：睡眠

図 12　名詞を付与した予測比を用いた関連度と被験者評価値の関係

度と被験者評価値の関係に名詞を付与した散布図を図 12 に示す．図 12 より，名詞の「肥満」や「食欲」では被験者評価値は中程度であるが，$log_{10}PR(q|n)$ が高くなっている．これは，これらの名詞が「睡眠」と同じ「健康」に関するものであるため，「健康には，十分な睡眠や肥満に気をつけた食生活が重要だ.」のように，「健康」の話題で同じ文書に出現する頻度が高くなり，被験者評価値と比較して $log_{10}PR(q|n)$ が高くなっていると考えられる．そのため，カテゴリーが同じ名詞との共起確率を用いた関連度は実際に人が感じる関連性と違いが大きい可能性があり，カテゴリーに関する調査や補正が必要である．

### 4.3　共起確率の予測における区間分割の評価

#### 4.3.1　評価方法

分割数を変化させると予測の共起確率が変化するため，予測比を用いた関連度も変化する．そこで，区間の分割数を変化させた際の予測比を用いた関連度と被験者評価値との関係を比較する．また各区間の $log_{10}P(q|n)$ の偏りを調べるため，$log_{10}|D_n|$ と $log_{10}P(q|n)$ の関係を密度ごとに色分けした二次元ヒストグラムを用いて調査を行う．

#### 4.3.2　結果と考察

3.2.1 項における分割数を 0〜10 で変化させ，$log_{10}|D_n|$ を等分割した際の予測比を用いた関連度と被験者評価値の各クエリの相関係数の平均の推移を図 13 に示す．また，区間の名詞数が予測精度に影響を与えると考え，各分割数での区間の名詞数の最小値を図 13 に合わせて示す．なお，分割数 0 は対数共起確率と被験者評価値との相関係数の平

© 2018 Information Processing Society of Japan

均値を表しており，区間の名詞数の最小値の軸は対数で表示している．図 13 より，相関係数の平均は分割数 2 以上では 0.59 付近でほとんど変化していないが，分割数を増加させるとわずかに減少することが確認できる．また，区間の名詞数の最小値は分割数を増加させると減少し，分割数 8 で名詞数が 100 を下回ることが確認できる．このことから，分割数の変化は予測値の変化に大きな影響を与えないが，分割数を増加させると各区間の名詞数が減少するため，予測値の正確な算出が難しくなり相関係数が減少していると考えられる．

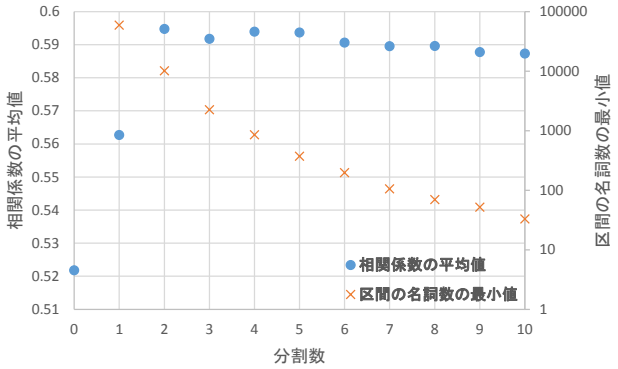
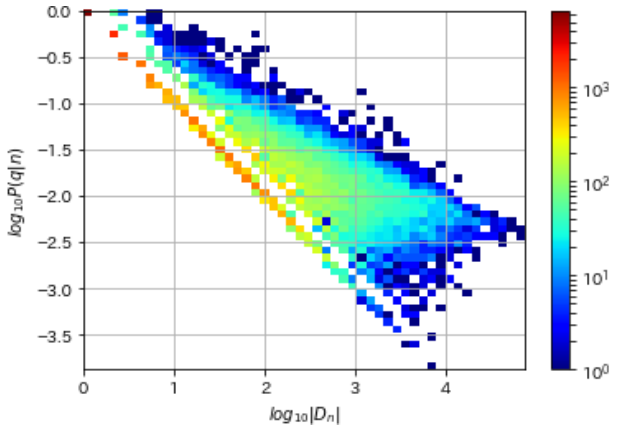



**図 13** 分割数と相関係数の平均，区間の名詞数の最小値の関係

次に，各区間の $log_{10}P(q|n)$ の偏りを調べるため，$log_{10}|D_n|$ と $log_{10}P(q|n)$ の関係をデータの密度によって色分けして調査を行う．クエリ「認知症」における $log_{10}|D_n|$ と $log_{10}P(q|n)$ の関係を密度ごとに色分けした二次元ヒストグラムを図 14 に示す．なお，密度として，$log_{10}|D_n|$ と $log_{10}P(q|n)$ を 50 等分した領域のデータの数を用いた．図 14 より，密度は連続的に変化していることが確認できる．また，$log_{10}|D_n|$ の高さにより，$log_{10}P(q|n)$ の分布の偏りが異なっていることが確認できる．そのため，今後の課題として，$log_{10}|D_n|$ の値による $log_{10}P(q|n)$ の分布の違いを考慮した，共起確率の予測法の開発が挙げられる．




クエリ：認知症

**図 14** 密度ごとに色分けした $log_{10}|D_n|$ と $log_{10}P(q|n)$ の関係

## 5. おわりに

本稿では，検索クエリに関する知識の無い人が Web 検索を行う際に信頼できる情報を選択する支援を行うために，語の組の関連性に注目し，関連度を共起確率の予測値に対する実測値の比を用いて算出する手法を提案した．その際，クエリと共起する全ての名詞の共起確率を名詞-文書データベースから算出し，$log_{10}|D_n|$ で分割した各区間の平均値をつないだ折れ線を予測の値とする手法を提案した．

また，クラウドソーシングを用いて関連度の被験者評価値を収集し，対数共起確率と提案手法の比較を行ったところ，20 クエリ中 15 クエリで相関係数が増加することが確認できた．$log_{10}|D_n|$ ごとに色分けした被験者評価値との関係を調べると，$log_{10}|D_n|$ による $log_{10}P(q|n)$ への影響を補正できていることが確認できた．

予測の共起確率を算出する際の $log_{10}|D_n|$ の分割数を変化させて予測比を用いた関連度と被験者評価値との相関係数を調べたところ，分割数による予測の共起確率の違いはほとんどないことが確認できた．

さらに，各区間の $log_{10}P(q|n)$ の偏りを調査するため，$log_{10}|D_n|$ と $log_{10}P(q|n)$ の関係を密度ごとに色分けしたところ，$log_{10}|D_n|$ の値による $log_{10}P(q|n)$ の分布の偏りの違いが確認できた．今後の課題として，分布の偏りを考慮した，共起確率の予測法を開発することが挙げられる．

**謝辞** 本研究の一部は，平成 30 年度科研費基盤研究 (C)(17K00429) によるものである．

### 参考文献

[1] 山本 祐輔，手塚 太郎，アダム ヤトフト，田中 克己：ページ特性を考慮した Web 検索結果の集約とページ生成時間分析による知識の信頼性判断支援，電子情報通信学会論文誌，Vol.J91-D，No.3，pp.576-584，2008.

[2] 山本 祐輔，田中 克己：反証センテンスの提示による信憑性指向のウェブ検索支援，情報処理学会論文誌：データベース， Vol.6，No.2，pp.42-50，2013.

[3] Akamine, S., Kawahara, D., Kato, Y., Nakagawa, T., Inui, K., Kurohashi, S., and Kidawara, Y.: WISDOM: A Web Information Credibility Analysis Systematic. In Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, Association for Computational Linguistics, pp. 14, 2009.

[4] Cilibrasi, Rudi L., and Paul MB Vitanyi.: The google similarity distance. IEEE Transactions on knowledge and data engineering 19.3, 2007.

[5] 渡部 広一，河岡 司：常識的判断のための概念間の関連度評価モデル，自然言語処理，Vol.8， No.2，pp.39-54，2001.

[6] 渡部 広一，奥村 紀之，河岡 司：概念の意味属性と共起情報を用いた関連度計算方式，自然言語処理，Vol.13，No.1，pp.53-74，2006.

[7] Kudo, T., Yamamoto, K., and Matsumoto, Y.: Applying conditional random fields to Japanese morphological analysis. Proceedings of the 2004 conference on empirical methods in natural language processing, pp. 230-237, 2004.

© 2018 Information Processing Society of Japan