Vol.2017-DBS-165 No.11
Vol.2017-IFAT-128 No.11
2017/9/18

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

小山 雄也[1]　湯本 高行[1]　礒川 悌次郎[1]　上浦 尚武[1]

概要：Web 検索において，信頼性に問題のある Web ページが推薦される問題がある．そのため，利用者は検索結果から信頼できる情報を自ら選択する必要がある．しかし，利用者が検索対象に関して知識がない場合，信頼性を判断することは困難であると考えられる．そこで本研究では，信頼性の判断を支援するために，典型性に注目する．なお，信頼性と典型性の関係として，典型性が高くなると，信頼性が高くなるとみなす．本研究では，文の典型性を定量化した典型度を文中の名詞の自己相互情報量 (PMI) を用いて表現する．PMI の算出において，クエリと文中の名詞の類義語を考慮することで表記揺れの影響を低減させる．また，文の典型度としきい値を比較することで典型性の推定を行う．そのため，典型性の推定に用いるしきい値の決定方法を提案する．

1. はじめに

近年，スマートフォンの普及により，Web 検索の利用者は増加し，誰でもどこでも行えるようになっている．Web 検索では利用者がクエリを入力すると，それに関係する Web サイトのランキングが表示され，そのランキングから利用者が自ら Web サイトを選択することで，情報が表示される．例えば，利用者の頭が痛い場合にクエリとして「頭痛」を入力し，検索すると頭痛をまとめたページのランキングが表示され，ある Web ページを選択すると，対処法や原因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の対処法などを知ることが可能である．一方で，Web 上には不正確で信憑性の低い情報も多数存在している．例えば，医療関係の Web サイトでも，医療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，利用者に知識がない場合は Web 情報の信憑性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，情報の典型性に注目し，クエリに関連する文の典型性の推定を行う．これを実現するために，文の典型性を定量化した典型度を文中の名詞の PMI を用いて表現する．なお，PMI を算出する際の語の表記揺れによる影響を低減するため，クエリと名詞の類義語を考慮している．その後，文の典型度としきい値を比較し，その文が典型的か非典型的かを出力する．

[1] 兵庫県立大学大学院工学研究科

2. 関連研究

語と語の関係を算出する研究として，岡らの研究 [1] がある．この研究では語と語の関係を，検索エンジンの検索結果の上位の Web ページでの語の出現数を用いて表している．本研究では，データベースを用いて関係性を算出している点で異なっている．

典型度の算出に関する先行研究として，山中らの研究 [2] がある．この手法では，クエリ q において文書 d が連想できない確率を $P(\bar{d}|q)$ とし，文書 d の全ての主要語 w_i がクエリ q と共起されない場合，文書 d はクエリ q に対して典型的でないと定義している．これより，非典型度は語 w_i がクエリ q から連想できない確率 $P(\bar{w_i}|q)$ の総積で算出可能であり，$P(\bar{w_i}|q)$ は，連想できる確率 $P(w_i|q)$ の余事象であるため，(1) 式で算出できる．

$$P(\bar{d}|q) = \prod_{w_i \in d} P(\bar{w_i}|q) = \prod_{w_i \in d} \{1 - P(w_i|q)\} \qquad (1)$$

また，(1) 式の $P(w_i|q)$ はクエリに対する語の共起確率を表し，(2) 式により算出する．(2) 式において，$|D_q|$ は DB において文書中にクエリ q を含む文書数を表し，$|D_q \cap D_{w_i}|$ は DB において文書中にクエリ q と語 w_i を同時に含む文書数を表す．

$$P(w_i|q) = \frac{|D_q \cap D_{w_i}|}{|D_q|} \qquad (2)$$

本研究では先行研究と異なり，文単位の典型度を用いるため，文章を対象に典型度を算出していた先行研究の手法を

用いるに当たって 2 つの問題点が存在する．

まず，1 つ目の問題点として，文の名詞を用いると文ごとに名詞数に違いがあるため，総積で典型度を算出すると名詞数が少ない語の値が高くなる傾向がある．そのため，本研究では値の総積ではなく，相加平均を行うことで名詞数による正規化を行い，典型度を算出する．

次に，2 つ目の問題点として，(2) 式の連想確率の算出では，一般的な語の連想確率が高くなってしまうことが挙げられる．先行研究では，典型度の算出に用いる語 w_i として文章中の名詞から主要語を抽出して用いている．本研究では文の典型度を算出するため，名詞数が少なく，一般的な語も典型度算出に用いている．そのため，クエリから語を連想できる確率として，クエリに対する語の共起確率ではなく，クエリと名詞の PMI を用いることで名詞自身が出現する文書数で正規化を行い，一般的な語の値が小さくなるように変更している．

そのほか，Web 検索の信頼性に関する先行研究として，山本らの研究がある [3][4]．[3] 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，[4] 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．これらの先行研究は，[3] はページ単位で信頼性を評価している点が，[4] は反証の文で Web の検索結果の注意喚起を行っている点が，それぞれ本研究とは異なっている．

3. 典型性推定手法

本研究ではユーザの入力した，クエリと文を用いて，クエリと文中の名詞との関係性から文単位の典型度を算出し，算出した典型度から文の典型性を推定することを目的とする．典型度算出の概略図を図 1 に示す．

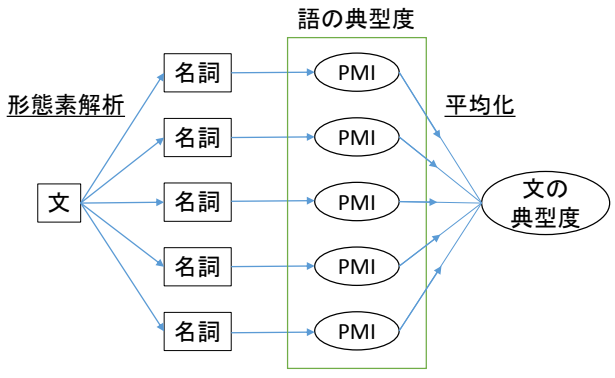

図 1 文の典型度算出

図 1 より，典型度算出の流れを以下に示す．

1 文を形態素解析し，名詞を抽出
2 語の出現確率を用いてクエリと名詞の PMI を算出
3 文の典型度として名詞の PMI の平均値を算出

2 の PMI 算出では，名詞-文書データベースにおける文書集合中の語の出現確率を用いる．また，出現確率を算出する際に，名詞とクエリの類義語を考慮することで表記揺れによる影響を低減している．

本研究では，算出された文の典型度をしきい値と比較することで典型性を推定する．

3.1 典型度算出

ある文 s があるクエリ q に対して連想できる指標を典型度 $T(q,s)$ とする．典型度 $T(q,s)$ はクエリ q と文書中の名詞 n_i の $PMI(n_i,q)$ の相加平均と定義すると，(3) 式で算出できる．

$$T(q,s) = \frac{1}{|s|}\sum_{n_i \in s} PMI(q,n_i) \tag{3}$$

なお，文書中の名詞 n_i がクエリ q である場合，その値は除去して計算する．ここで，(3) 式の $|s|$ は文に含まれる名詞の数を表す．また，文書中でのクエリ q の出現確率を $P(q)$，名詞 n の出現確率を $P(n)$，クエリ q と名詞 n の共起確率を $P(q,n)$ とおくと，$PMI(q,n_i)$ は (4) 式で定義される．

$$PMI(q,n_i) = \log_2 \frac{P(q,n_i)}{P(q)P(n_i)} \tag{4}$$

ここで，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 $|D|$ の商と定義すると，(4) 式は (5) 式のように変形できる．

$$PMI(q,n_i) = \log_2 \frac{|D||D_q \cap D_{n_i}|}{|D_q||D_{n_i}|} \tag{5}$$

なお，共起文書数 $|D_q \cap D_{n_i}|$ が 0 の場合は，$PMI(q,n_i) = 0$ としている．

3.2 類義語抽出

クエリと文中の名詞のみを用いた場合，表記揺れで PMI の値が大きく変化することがある．そこで，クエリと名詞の類義語を考慮することで，表記揺れを低減させる．類義語の抽出には Wikipedia[*1] 情報を保存したデータベースを使用し，Wikipedia の機能であるリダイレクトおよびリンクアンカーの関係を用いる．リダイレクトは，記事名の略称や別名および別の表記で記事を検索しても，実際の記事へ転送されるようにする機能であり，語と語の関係を表している．リンクアンカーは，記事の本文の語に，その語の記事へのリンクをつけることで，本文中の語を詳しく調べることができる機能であり，語とページの関係を表している．

類義語は最大で 10 語となるように抽出する．はじめに，ある語がリダイレクトする語を類義語として抽出する．抽出された語数が 10 語未満の場合，同じリンク先を持つ語

[*1] https://ja.wikipedia.org/

で，リンク数が上位の語を類義語が 10 語になるまで抽出している．

ある語 w の類義語を $S(w)$ とおくと (5) 式における各文書数は (6)，(7)，(8) 式のようになる．

$$D_q = \bigcup_{q_i \in S(q)} D_{q_i} \tag{6}$$

$$D_{n_i} = \bigcup_{n_{i_j} \in S(n_i)} D_{n_{i_j}} \tag{7}$$

$$D_q \cap D_{n_i} = \bigcup_{q_i \in S(q), n_{i_j} \in S(n_i)} D_{q_i} \cap D_{n_{i_j}} \tag{8}$$

3.3 典型性推定

典型性の推定には，3.1 節の手法を用いて算出された典型度を用いる．しきい値を η，典型度を $T(q,s)$ とおいたとき，(9) 式を満たす場合に典型的な文と推定する．

$$T(q,s) > \eta \tag{9}$$

典型度としきい値の関係から典型性を推定する際の問題としてクエリごとに典型度の値が大きく異なることが挙げられる．そのため，しきい値をクエリごとに推定する必要がある．そこで，η をクエリ q の関数 $\eta(q)$ と考えると，(9) 式は (10) 式に書き換えられる．

$$T(q,s) > \eta(q) \tag{10}$$

ここで，典型度の算出に用いる 3.1 節の PMI の式を変形した式を (11) 式に示す．

$$PMI(q,n_i) = \log_2 \frac{|D|}{|D_q|} + \log_2 \frac{|D_q \cap D_{n_i}|}{|D_{n_i}|} \tag{11}$$

(11) 式より，$\log_2 \frac{|D|}{|D_q|}$ はクエリごとの定数となる．そこで，クエリごとの基準値 $Ref(q)$ を (12) 式と定義する．

$$Ref(q) = \log_2 \frac{|D|}{|D_q|} \tag{12}$$

また，$\log_2 \frac{|D_q \cap D_{n_i}|}{|D_{n_i}|} \leq 0$ より，PMI は $Ref(q)$ を最大値にとり減少する．そこで，しきい値 $\eta(q)$ をクエリごとの基準値 $Ref(q)$ を用いて (13) 式と定義する．

$$\eta(q) = Ref(q) \times p \tag{13}$$

p は $[0,1]$ であり，基準値 $Ref(q)$ に対するしきい値 $\eta(q)$ の割合を表している．(13) 式を (10) 式に代入すると，典型性推定の式は (14) 式と表せる．

$$T(q,s) > Ref(q) \times p \tag{14}$$

これらより，(14) 式を満たす場合は典型的な文，満たさない場合は非典型的な文と推定する．

しきい値の取り方として以下が考えられる．

(a) クエリに依存しない，典型度の平均を用いたしきい値

表 1 評価実験に使用した入力クエリ

水素水	プラシーボ	ダイエット	睡眠
勉強法	集団的自衛権	アベノミクス	
EU	オリンピック	ポケモン	

表 2 構築した名詞-文書データベースの規模

URL 数	158,999
原形名詞数	2,065,821
出現形名詞数	44,999
名詞-URL の対応関係数	26,251,439

(b) クエリに依存する，固定の割合を用いたしきい値

(a) は，クエリごとの典型度の差を考慮しないしきい値を用いた典型性推定であり，しきい値として，各クエリの典型的な文と非典型的な文の典型度の平均値を平均した値を典型性推定を行う．(b) は，クエリごとの基準値 $Ref(q)$ からの一定の割合をしきい値とした場合の典型性推定を行う．以降，(a) をクエリ非依存のしきい値，(b) をクエリ依存のしきい値と表記する．

4. 評価実験

本研究ではクエリに対する典型度を算出し，評価実験では入力に表 1 のクエリを用いた．クエリには，身近な事柄であり，専門知識がなければ典型性の判断が難しいと推測される医療，健康に関する単語や，生活，政治，アニメゲームに関する単語を 10 語用いている．

次に，本研究で典型度の算出に用いた名詞-文書データベースの規模を表 2 に示す．このデータベースは，2016 年 8 月 12 日から 2016 年 10 月 31 日までの期間のはてなブックマーク[*2]のホットエントリーから構築したものである．

また，文単位の典型度を算出するために，それぞれのクエリに関する典型的な文と非典型的な文をあらかじめ用意する．評価実験に用いたクエリに関する典型的な文と非典型的な文の一部を表 3 に示す．典型的な文には，クエリの意味や，扱い方を説明した文を用いている．また，非典型的な文には，クエリに関する詳細な情報を述べている文を用いる．なお，文の選択は，著者のうちの一名が行った．

本研究では，文から名詞を抽出するために形態素解析を行っている．形態素解析器には MeCab[*3][5] を用いており，形態素解析辞書には mecab-ipadic-neologd[*4](neologd) を用いる．

4.1 PMI と共起確率に基づ典型度の比較実験

4.1.1 評価方法

本研究の手法である，PMI に基づく典型度に対して，先

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

表 3　入力文の一部

クエリ	典型的な文
水素水	水素分子のガスを溶解させた水であり，無味，無臭，無色である。
プラシーボ	実際には効果のないはずの治療を施すことによってよい効果が現れること
ダイエット	太りすぎを防ぐために低カロリーの食品をとること
クエリ	非典型的な文
水素水	水素は常温下においてはフッ素としか反応しない
プラシーボ	プラセボの色や数がプラセボ効果の方向性や大きさに影響する
ダイエット	マヨネーズは高脂肪食だが低ＧＩ食品であり『太らない』

行研究の手法である，クエリに対する名詞の共起確率に基づく典型度の比較を行う．本研究の手法が先行研究の手法と異なる点として，クエリと名詞の関係性を算出する際に，名詞の出現文書数で正規化を行うことが挙げられる．そこで，3.1 節で提案した典型度算出手法の，(5) 式における，名詞の出現文書数による正規化の有無で比較を行う．

正規化を行わない場合のクエリ q と文の名詞 n_i の関係性 $R(q, n_i)$ を (15) 式に示す．

$$R(q, n_i) = \log_2 \frac{|D||D_q \cap D_{n_i}|}{|D_q|} \tag{15}$$

(3) 式における，$PMI(q, n_i)$ を (15) 式の $R(q, n_i)$ とした際の値の変化を比較する．なお，名詞の出現文書数による正規化の影響を評価するため，本研究のもう一つの特徴である，類義語による拡張は行わない．２つの手法間で算出される典型度の差が大きいため，値の直接的な比較ではなく，クエリの総数に対する，典型的な文の典型度が非典型的な文の典型度を上回っているクエリ数の割合を２値判別率とし，その値で評価を行う．

4.1.2　結果と考察

$PMI(q, n_i)$ を用いて算出した結果および，(15) 式を用いて算出した結果を表 4 に示す．表 4 より，２値判別率が向上していることが確認できる．

また，表 4 より，PMI を用いることによって，２つのクエリの２値判別が可能になり，１つのクエリの判別が不能になっていることが確認できた．そこで，２値判別が不能になった「プラシーボ」に関して，典型的な文と非典型的な文に出現する名詞の出現数，クエリと名詞の共起数，$R(q, n_i)$ および，PMI を表 5 に示す．表 5 より，２値判別が不能になった原因として，非典型的な文に出現する「プラセボ」及び，「プラセボ効果」の名詞の出現数が極端に少ないため PMI が大きくなっているためであると推測できる．これらの名詞は，「プラシーボ」の類義語であると推測できるため，類義語を考慮し，クエリの類義語を典型度算出から除くことで対策可能である．

表 4　正規化による２値判別の変化

	PMI			正規化なし		
クエリ	典型	非典型	判別	典型	非典型	判別
水素水	4.943	4.948	×	12.68	13.89	×
プラシーボ	2.960	3.754	×	15.45	13.73	○
ダイエット	4.362	4.108	○	14.79	12.20	○
睡眠	2.707	2.561	○	13.59	14.18	×
勉強法	3.259	1.440	○	15.59	14.78	○
集団的自衛権	4.239	0.000	○	14.38	0.000	○
アベノミクス	4.583	3.161	○	13.28	14.64	×
EU	2.563	0.907	○	10.94	10.45	○
オリンピック	2.620	2.178	○	14.30	12.90	○
ポケモン	3.516	1.843	○	13.61	12.64	○
判別率	80%			70%		

表 5　「プラシーボ」(出現数：48) に関する文の名詞

	名詞	出現数	共起数	$R(q, n_i)$	PMI
典型的な文	効果	11,921	26	16.39	2.85
	治療	2,765	7	14.50	3.07
非典型的な文	影響	14,406	16	15.69	1.88
	数	8,406	2	12.69	-0.34
	色	6,982	5	14.02	1.25
	方向性	1,769	2	12.69	1.90
	プラセボ	39	5	14.02	8.73
	プラセボ効果	18	3	13.28	9.11

これらより，名詞の出現数による正規化は，極端に名詞の出現数が少ない場合を除いて有効であることが推測できる．そのため，以降の実験では PMI を用いた手法で典型度を算出していく．

4.2　類義語を考慮した典型度の評価実験

4.2.1　評価方法

本研究の提案手法として，類義語を考慮した典型度の算出が挙げられる．そこで，類義語を考慮する場合と考慮しない場合の典型度を評価する．類義語を考慮する手法として，3.2 節における類義語拡張した文書数を，3.1 節の PMI の算出に適用した場合の典型度算出を用いる．類義語拡張による影響を評価するため，4.1 節と同様に２値判別率を評価する．

4.2.2　結果と考察

類義語を考慮した典型度と２値判別および，考慮しない場合の典型度と２値判別を表 6 に示す．表 6 より，２値判別率が向上していることが確認できた．

表 6 より，類義語を考慮することによって，２つのクエリの２値判別が可能になり，１つのクエリの判別が不能になっていることが確認できた．そこで，２値判別が可能になった「プラシーボ」に関して，類義語を考慮した場合の典型的な文と非典型的な文に出現する名詞の出現数およ

表 6 類義語考慮の有無による 2 値判別比較

	類義語を考慮			類義語を考慮しない		
クエリ	典型	非典型	判別	典型	非典型	判別
水素水	4.052	2.242	○	4.943	4.948	×
プラシーボ	3.393	1.294	○	2.960	3.754	×
ダイエット	2.396	2.471	×	4.362	4.108	○
睡眠	1.861	1.840	○	2.707	2.561	○
勉強法	2.423	1.357	○	3.259	1.440	○
集団的自衛権	3.082	0.000	○	4.239	0.000	○
アベノミクス	3.453	2.673	○	4.583	3.161	○
EU	2.921	1.661	○	2.563	0.907	○
オリンピック	2.022	1.395	○	2.620	2.178	○
ポケモン	3.451	0.966	○	3.516	1.843	○
判別率	90%			80%		

表 7 「プラシーボ」(出現数：128) に関する文の名詞

	名詞	出現数	共起数	PMI
典型的な文	効果	20,304	107	2.71
	治療	3,463	47	4.08
非典型的な文	色	10,369	20	1.26
	数	22,057	36	1.02
	方向性	1,769	3	1.07
	影響	16,505	47	1.82

び，クエリと名詞の共起数を表 7 に示す．表 7 より，4.1.2 節で挙げた，2 値判別ができない原因となっていた名詞の「プラセボ」および，「プラセボ効果」がクエリの類義語と判断され，典型度算出から除去されていることが確認できた．これにより，非典型的な文の典型度が減少することで，判別が可能になったと推測できる．

また，判別が不能になった「ダイエット」に関して，類義語を考慮する場合としない場合における，それぞれの典型的な文と非典型的な文に出現する名詞の出現数および，クエリと名詞の共起数を表 8 に示す．表 8 より，判別が不能になった原因として，非典型的な文の名詞「GI」の出現数，共起数ともに変化しなかったことが挙げられる．これによって，「GI」の PMI の減少が他のクエリに対して小さくなり，典型的な文の典型度が非典型的な文の典型度を下回ることになった．名詞「GI」の特徴として，出現数共起数ともに，他の名詞に比べて少ないことが確認できる．そこで，出現数，共起数が極端に少ない名詞を除去するなどの対策が必要である．

4.3 典型性推定におけるしきい値の決定方法の比較実験

4.3.1 評価方法

3.3 節の推定手法および，2 つのしきい値を用いた典型性推定を比較する．ここで，4.2 節の表 6 より，各クエリの典型的な文と非典型的な文の類義語を考慮した典型度の平均値を求めた結果，3.3 節の (a) のクエリ依存のしきい値は，$\eta = 2.248$ となった．また，(b) のクエリごとの基準値からの割合を最大値の 4 分の 1 である，$p = 25\%$ として比較を行う．

表 8 「ダイエット」に関する文の名詞の比較

	類義語を考慮			類義語を考慮しない		
名詞	出現数	共起数	PMI	出現数	共起数	PMI
ダイエット	8,261			1,852		
カロリー	5,537	771	2.64	1,151	436	5.02
食品	6,815	680	2.16	1,658	251	3.70
マヨネーズ	745	83	2.31	446	52	3.32
脂肪	6,693	715	2.25	751	346	5.31
食品	6,815	680	2.16	1,658	251	3.70
ＧＩ	10	2	3.16	10	2	4.10

表 9 しきい値の取り方の違いによる典型性推定の比較

	クエリ非依存のしきい値		クエリ依存のしきい値	
クエリ	典型	非典型	典型	非典型
水素水	T	A	T	A
プラシーボ	T	A	T	A
ダイエット	T	T	T	T
睡眠	A	A	T	T
勉強法	T	A	T	A
集団的自衛権	T	A	T	A
アベノミクス	T	T	T	T
EU	T	A	T	T
オリンピック	A	A	T	A
ポケモン	T	A	T	A

典型性推定の評価指標として，適合率，再現率，F 値を用いる．適合率は典型的と推定された文のうち実際に典型的な文をどれだけ含んでいるかという正確性の指標であり，再現率は推定対象としている文の中で，正解文のうちでどれだけの文を推定できているかという網羅性の指標である．また，F 値は適合率と再現率の調和平均であり，特徴として，適合率と再現率がともに高い場合に高い値を示す．そのため，F 値を比較することで典型性推定の精度を評価する．

4.3.2 結果と考察

2 つのしきい値の典型性推定の結果を表 9 に，適合率，再現率，F 値を算出した結果を表 10 に示す．なお，典型的な文と推定した場合を T，非典型的な文と推定した場合を A としている．表 9 より，典型性推定の F 値が高い推定手法はクエリ依存のしきい値であり，$p = 25\%$ としたときであることが確認できた．この結果から，しきい値を用いた典型度推定には，クエリごとの典型度の差を考慮することが有効であると考えられる．

5. 分析・調査

5.1 PMI の出現数特性

4.1.2 および，4.2.2 より，名詞の出現数が極端に少ない

表 10　適合率，再現率，F 値の比較

	クエリ非依存のしきい値	クエリ依存のしきい値
適合率	0.800	0.714
再現率	0.800	1.000
F 値	0.800	0.833

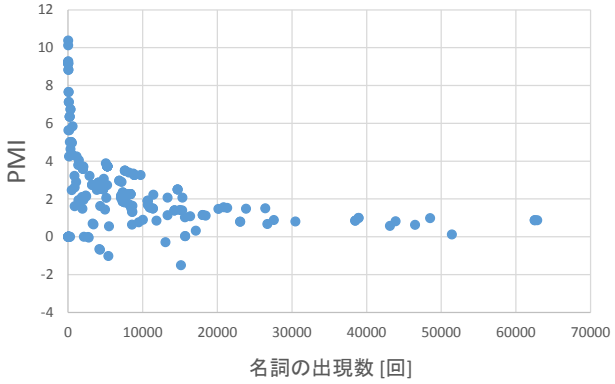

図 2　「水素水」における出現数-PMI 特性

場合，PMI 及び典型度が正しく算出できないと推測された．そこで，名詞の出現数と PMI の関係を調査する．

Wikipedia の「水素水」に関する記事の先頭 5 段落の 19 文を抽出し，PMI を求めた際の名詞の出現数-PMI 特性を図 2 に示す．図 2 より，頻度が少なくなるとばらつきが大きくなることが確認できた．このことから，低頻度の語の PMI に対して補正を行う必要があると考える．

5.2　共起を用いたクエリと名詞の関係性の分析

4.2 節の表 7 において，典型的な文の「治療」の PMI が「効果」に比べてかなり大きいことが確認できる．このように，PMI は人での評価では関連度に違いのなさそうな 2 語の値の差が大きくなる場合がある．そこで，クエリと名詞の両方に共起する語の関連度のランキングから上位の語を抽出することで，クエリと名詞の関係性を分析する．

5.2.1　共起語の抽出 (ランキング)

ある語の組に共起する文書に共起する語のうち，PMI のランキングの上位を抽出し，語の組の関係性を分析する．そこで，ある語 w_1 と語 w_2 の共起する文書に共起する語を共起語 c とすると，w_1 と w_2 の組 (w_1, w_2) と c_i との関連性 $PMI(w_1 \wedge w_2, c_i)$ はクエリと名詞の PMI の (5) 式を変形した (16) 式で表せる．

$$PMI(w_1 \wedge w_2, c_i) = \log_2 \frac{|D||D_{w_1 \wedge w_2} \cap D_{c_i}|}{|D_{w_1 \wedge w_2}||D_{c_i}|} \qquad (16)$$

(16) 式を用いて，「鳥」と「旅客機」の文脈語の上位 7 語を表 11(a) に「木」と「本」の文脈語の上位 7 語を表 11(b) に示す．表 11 より，「鳥」と「旅客機」に関しては関連する語が抽出されていることがわかる．一方で，「木」と「本」を入力した場合，ほとんど関係のない語が抽出されている．このことから，共起語は語の組の関係性を間接的に表している可能性がある．これらを用いることで，共起語を用いた語の組の関係性の定量化が期待できる．

表 11　(a)「鳥」と「旅客機」　(b)「木」と「本」

名詞	PMI
着陸	7.443
乗客	7.280
墜落	7.236
フライト	7.072
機体	6.989
戦闘機	6.901
航空	6.860

名詞	PMI
スレ	3.665
マジ	3.578
いや	3.539
わからん	3.523
馬鹿	3.505
アホ	3.504
ゴミ	3.481

6.　おわりに

本稿では，検索結果から信頼できる情報を選択する支援を行うために典型性に注目し，典型性を文単位で推定する手法を提案した．また，名詞の PMI と出現数の関係を調査し，共起語を用いたクエリと名詞の関係性の分析を行った．

先行研究で用いられていた関係性を，PMI を用いた典型度に変更した結果，判別率が向上した．また，類義語の考慮を比較した結果も同様に判別率が向上した．

さらに，典型性推定のためのしきい値の決定方法を提案し，2 つの決定方法で比較実験を行った結果，典型性推定にはクエリ依存のしきい値を用いると精度が良いことが確認できた．

名詞の PMI と出現数の関係調査では，低頻度語が PMI に与える影響がほかの語より大きいことが確認できた．そのため，低頻度語の PMI の補正が必要である．また，共起語を用いた語の組の関係性の分析では，語の組と共起する語の PMI のランキングの上位が関係性を間接的に表していることが確認できた．これを用いて，共起語から語の組の関係性を定量的に算出することが期待できる．

謝辞　本研究の一部は，平成 29 年度科研費基盤研究 (C)(17K00429) によるものである．

参考文献

[1] 岡 瑞起，松尾 豊：検索エンジンを用いた関係の重みづけ，人工知能学会論文誌，Vol.25，No.1，pp.1-8，2010.
[2] 山中 隆広，湯本 高行，新居 学，上浦 尚武：ページとクエリの連想関係に基づく希少な Web ページの検索，WebDB Forum 2015，A2-3，2015.
[3] 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.
[4] 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.
[5] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237, 2004.