Web 文書の抽象度および具体度の測定にむけて

Towards Measuring Concreteness and Abstractness Levels of Web Documents

田中 伸弥 † アダム ヤトフト † 田中 克己 †
Shinya Tanaka Adam Jatowt Katsumi Tanaka

1. はじめに

現在，Web 上には多くの Web 文書が存在している．それらの Web 文書の中には，抽象的に記述されたものと，具体的に記述されたものとが存在する．

具体性とは，はっきりとした実態を備えている事であり，人間の五感で感じることができる事物であることである [1]．また，心的イメージの形成のしやすさを用いて具体性を示す場合もある [2]．

具体的に記述された文書は，読者のイメージの形成を助け，読者の理解を助ける [2]．

近年インターネットから情報を取得するために Web 検索エンジンが頻繁に利用されている．既存の Web 検索エンジンはユーザの入力したクエリに対して，適合度やリンクによる支持度等によってランキングされた Web ページのリストを出力する．得られた検索結果のリストの中には，入力クエリの抽象的な記述を含む Web ページと，入力クエリの具体的な記述を含む Web ページが混在している．このため検索結果のリストの中から具体的な記述，もしくは抽象的な記述を含む Web ページを探すのは容易ではない．

本稿では，具体的な記述，もしくは抽象的な記述を含む Web ページの検索を行うために，Web 文書の抽象度及び具体度の測定手法を提案する．Web 検索結果の Web ページのリストを，Web 文書の抽象度及び具体度に応じて並び替えることにより，検索結果のリストの中から具体的な記述，もしくは抽象的な記述を含む Web ページを容易に見つけ出すことができると考えられる．

本稿では，まず，単語レベルでの抽象性及び具体性に焦点をあてる．その後，Web 文書の抽象度及び具体度を求める手法を提案する．

2. 関連研究

加藤ら [3] は抽象的な入力クエリ X に対して，「X らしい」画像集合を検索する手法を提案した．加藤らは，抽象的な概念を複数の語集合で表現することにより，抽象的な概念の具体化を図った．

中村ら [4] は Web 検索結果のページ中に含まれる，味覚や嗅覚，視覚，聴覚など各種の感覚情報に着目し，感覚情報の抽出手法や感覚の可視化手法，感覚情報に基づくリランキング手法を提案した．

心理言語学においては，Paivio ら [1] は単語の具象性 (concreteness) と心像性 (imageability) を定義した．具象性とは視覚や聴覚などの感覚によって経験することができる度合いを示す属性であり，心像性とは心的イメージの思い浮かべやすさの度合いを示す属性である．1 節で示した具体性の観点より，具象性及び心像性の値が単語の抽象度及び具体度を表すと考えられる．

3. 具象性および心像性

3.1. 概要

Web 文書の抽象度及び具体度を測定するために，本稿では，まず，単語レベルでの抽象度及び具体度に焦点をあてる．

Paivio ら [1] の用いた心理言語学における，単語の持つ具象性の値 (具象性スコア) 及び心像性の値 (心像性スコア) を，単語の抽象度及び具体度を表す指標として用いる．

本節では，具象性スコア及び心像性スコアと，Web 検索エンジンと画像検索エンジンにおける検索結果数，概念階層における単語の位置及びソーシャルタギング情報との相関を示す．

単語の具象性スコア及び心像性スコアには MRC Psycholinguistic Database (MRCDB)[5] の値を用いた．

表 1, 2 に本節で用いた MRCDB における具象性スコア及び心像性スコアを示す．MRCDB における具象性スコア及び心像性スコアの理論上の最小値は 100，最大値は 700 である．

表 1: MRCDB における具象性スコア

品詞	単語数	最小値	最大値	平均値	標準偏差
名詞	3479	158	670	458.5	117.3
動詞	1458	194	670	448.9	110.0
形容詞	493	183	622	371.8	87.5
その他	169	158	639	311.6	95.4
すべて	5599	158	670	443.9	117.4

表 2: MRCDB における心像性スコア

品詞	単語数	最小値	最大値	平均値	標準偏差
名詞	3479	129	667	471.6	105.4
動詞	1458	202	667	473.5	95.9
形容詞	493	129	619	425.6	83.5
その他	169	143	630	341.3	107.2
すべて	5599	129	667	461.1	104.4

表 3: 具象性スコアと心像性スコアの相関係数

品詞	相関係数
名詞	0.83
動詞	0.86
形容詞	0.77
その他	0.86
すべて	0.84

表 3 に具象性スコアと心像性スコアの相関係数を示す．具象性スコアと心像性スコアには高い相関がみられる．相関を示さない例として，具象性スコアが高く，心像性スコアが低い単語として，astrolabe，coffer などが挙げられる．これらの単語は使用頻度く，心的イメージが想起しにくいため具象性スコアと心像性スコアの差が大きいと考えられる．一方，具象性スコアが低く，心像性スコアが高い単語として，devil，infinity などが挙げられる．これらの単語の表す概念自体には物理的な形状が無いが，単語を表すシンボルなどが存在する．そのため心的イメージを想起しやすく，具象性スコアと心像性スコアの差が大きいと考えられる．

3.2. 検索結果数との相関

具体性は心的イメージの形成のしやすさと関連があると考えられるため，Web 検索エンジンにおける検索結果数と画像検索エンジンにおける検索結果数との相関を調査した．単語 t に対して，t の Web 検索エンジンにおける検索結果数を $h_{\mathrm{web}}(t)$ とし，t の画像検索エンジンにおける検索結果数を $h_{\mathrm{img}}(t)$ とする．$h_{\mathrm{web}}(t)$，$h_{\mathrm{img}}(t)$ の対数を取った値を，$h_{\mathrm{weblog}}(t) = \log_{10}(h_{\mathrm{web}}(t)+1)$，$h_{\mathrm{imglog}}(t) = \log_{10}(h_{\mathrm{img}}(t)+1)$ とする．また，これらの比を，$r_{\mathrm{hit}}(t) = \frac{h_{\mathrm{img}}(t)+1}{h_{\mathrm{web}}(t)+1}$，$r_{\mathrm{hitlog}}(t) = \frac{h_{\mathrm{imglog}}(t)}{h_{\mathrm{weblog}}(t)}$ とする．

Web 検索エンジンには Bing[1]を，画像検索エンジンには Flickr[2]を用いた．

表 4，表 5 に具象性スコア及び心像性スコアと検索結果数との相関係数を示す．具象性スコア，心像性スコア共に画像検索エンジンでの検索結果数 $h_{\mathrm{imglog}}(t)$ との相関係数が高い事がわかり，Web 検索エンジンでの検索結果数 $h_{\mathrm{weblog}}(t)$ を用いることで相関係数が増加する事がわかる．

3.3. 概念階層における深さとの相関

概念間の上位下位関係や部分全体関係などの意味的な階層関係と具象性スコア及び心像性スコアの相関を調査した．

単語 t は通常複数の意味を持ち，それぞれの意味で概念階層における位置が異なる．根ノードから t の最も使用頻度の高

†京都大学情報学研究科社会情報学専攻

[1] http://www.bing.com
[2] http://www.flickr.com

表 4: 具象性スコアと検索結果数の相関係数

品詞	相関係数					
	h_{web}	h_{img}	r_{hit}	h_{weblog}	h_{imglog}	r_{hitlog}
名詞	-0.09	0.13	0.23	0.06	0.42	0.53
動詞	-0.20	0.15	0.25	-0.02	0.45	0.58
形容詞	-0.11	0.20	0.22	0.02	0.32	0.41
その他	-0.30	0.26	0.32	-0.32	0.17	0.40
すべて	-0.23	0.14	0.24	-0.06	0.37	0.53

表 5: 心像性スコアと検索結果数の相関係数

品詞	相関係数					
	h_{web}	h_{img}	r_{hit}	h_{weblog}	h_{imglog}	r_{hitlog}
名詞	-0.03	0.22	0.25	0.23	0.57	0.62
動詞	-0.19	0.23	0.31	0.03	0.53	0.65
形容詞	-0.11	0.27	0.20	0.10	0.46	0.54
その他	-0.26	0.26	0.30	-0.20	0.24	0.41
すべて	-0.20	0.22	0.26	0.06	0.49	0.61

い意味における位置までの深さを $d_{\mathrm{freq}}(t)$ とする．また，根ノードから t の取りうる位置までの深さのうち，最小値，最大値及び平均値を $d_{\mathrm{min}}(t)$，$d_{\mathrm{max}}(t)$，$d_{\mathrm{avg}}(t)$ とする．

実験には WordNet[3] の概念階層を用いた．WordNet で利用できる概念階層が名詞と動詞のみであるため，本実験は名詞と動詞のみを用いて行った．

表 6: 具象性スコアと概念階層における深さの相関係数

品詞	相関係数				
	意味数	d_{freq}	d_{min}	d_{max}	d_{avg}
名詞	0.01	0.31	0.27	0.24	0.31
動詞	-0.18	0.10	-0.10	0.15	0.08
すべて	-0.07	0.20	0.15	0.17	0.18

表 7: 心像性スコアと概念階層における深さの相関係数

品詞	相関係数				
	意味数	d_{freq}	d_{min}	d_{max}	d_{avg}
名詞	0.08	0.23	0.24	0.13	0.23
動詞	-0.17	0.09	-0.10	0.12	0.07
すべて	-0.03	0.13	0.10	0.08	0.11

表 6，表 7 に具象性スコア及び心像性スコアと，概念階層における深さの相関係数を示す．名詞の具象性スコアと概念階層における根ノードからの距離とは相関がある事がわかる．心像性スコアにおいては，具象性スコアほどの相関は無い事がわかる．動詞に関して，名詞ほどの相関を示さない原因として，WordNet における名詞の概念階層と動詞の概念階層の構造の違いが影響していると考えられる．WordNet での名詞の概念階層は，単一の共通した根ノードを持つが，動詞の概念階層では単一の共通した根ノードを持たないことが影響していると考えられる．

3.4. 画像のソーシャルタグ数との相関

ある画像から連想される単語と，単語の抽象度及び具体度との関連を調査するため，画像検索エンジンのソーシャルタグ数と，単語の具象性スコア及び心像性スコアとの相関を調査した．

単語 t に対して，t の画像検索結果の画像に対して付けられたタグの総数を $n_{\mathrm{tag}}(t)$ とし，タグの種類数を $n_{\mathrm{uniq}}(t)$ とする．$n_{\mathrm{tag}}(t)$，$n_{\mathrm{uniq}}(t)$ の対数を取った値を，$n_{\mathrm{taglog}}(t) = \log_{10}(n_{\mathrm{tag}}(t)+1)$，$n_{\mathrm{uniqlog}}(t) = \log_{10}(n_{\mathrm{uniq}}(t)+1)$ とする．また，これらの比を $r_{\mathrm{tag}}(t) = \frac{n_{\mathrm{uniq}}(t)+1}{n_{\mathrm{tag}}(t)+1}$，$r_{\mathrm{taglog}}(t) = \frac{n_{\mathrm{uniqlog}}(t)}{n_{\mathrm{taglog}}(t)}$ とする．

画像検索エンジンには Flickr を用いて，単語 t でタグ検索した検索結果のうち，上位 1000 件までのソーシャルタグ情報を用いた．

表 8，表 9 に具象性スコア及び心像性スコアとソーシャルタグ数との相関係数を示す．具象性スコア及び心像性スコアとソーシャルタグの数に対する種類数 $r_{\mathrm{taglog}}(t)$ との相関があり，より多様なタグが付けられるほど具象性スコア及び心像性スコアが増加すると考えられる．

[3]http://wordnet.princeton.edu

表 8: 具象性スコアとソーシャルタグ数の相関係数

品詞	相関係数					
	n_{tag}	n_{uniq}	r_{tag}	n_{taglog}	n_{uniqlog}	r_{taglog}
名詞	0.08	0.29	0.20	0.22	0.34	0.35
動詞	0.08	0.33	0.27	0.19	0.35	0.38
形容詞	0.08	0.26	0.16	0.13	0.26	0.26
その他	-0.07	0.09	0.17	-0.06	0.04	0.09
すべて	0.04	0.24	0.19	0.17	0.28	0.30

表 9: 心像性スコアとソーシャルタグ数の相関係数

品詞	相関係数					
	n_{tag}	n_{uniq}	r_{tag}	n_{taglog}	n_{uniqlog}	r_{taglog}
名詞	0.22	0.48	0.27	0.35	0.52	0.52
動詞	0.16	0.43	0.30	0.27	0.45	0.48
形容詞	0.21	0.44	0.17	0.30	0.45	0.40
その他	0.05	0.18	0.19	0.11	0.18	0.20
すべて	0.16	0.40	0.25	0.28	0.44	0.45

4.Web 文書の抽象度及び具体度

3 節の結果より，単語 t に対して，Web 検索エンジンでの検索結果数に対する画像検索エンジンでの検索結果数 $r_{\mathrm{hitlog}}(t)$，概念階層における根ノードからの平均の深さ $d_{\mathrm{avg}}(t)$，ソーシャルタグ数に対するソーシャルタグの種類数 $r_{\mathrm{taglog}}(t)$ と，t の具象性スコア及び心像性スコアとの相関が高い事がわかる．これらを用いて単語 t の具体度 $Conc_{\mathrm{term}}(t)$，及び Web 文書 D を単語の集まりとしたとき，D の具体度 $Conc_{page}(D)$ を以下のように提案する．

$$Conc_{\mathrm{term}}(t) = \alpha \cdot r_{\mathrm{hitlog}}(t) + \beta \cdot d_{\mathrm{avg}}(t) + \gamma \cdot r_{\mathrm{taglog}}(t), \quad (1)$$

$$Conc_{\mathrm{page}}(D) = \frac{1}{|D|}\sum_{t \in D} Conc_{\mathrm{term}}(t) \quad (2)$$

ここで，$\alpha+\beta+\gamma=1$ とする．

5. まとめと今後の課題

本稿では，Web 文書の抽象度及び具体度の測定にむけた，心理言語学における具象性スコア及び心像性スコアと，Web 検索エンジンと画像検索エンジンにおける検索結果，概念階層での単語の位置及びソーシャルタギング情報との相関を調査した．調査の結果を踏まえて，単語の具体度と Web 文書の具体度を，(1) 式と (2) 式で提案した．

今後は (1) 式における最適な係数を求め，Web ページの抽象度及び具体度と，単語レベルでの抽象度及び具体度の関係の調査を行う予定である．

参考文献

[1] Paivio A, Yuille JC, and Madigan SA. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology: Monograph Supplement*, Vol. 76, No. 1, pp. 1–24, 1968.

[2] John Friedlander. Abstract, concrete, general, and specific terms. `http://grammar.ccc.commnet.edu/grammar/composition/abstract.htm`.

[3] 加藤誠, 大島裕明, 小山聡, 田中克己. Web 画像の「らしさ」検索: 語の典型的特徴を表す語集合のソーシャルタギング情報からの取得による web 画像検索. 電子情報通信学会第 19 回データ工学ワークショップ (DEWS2008), March 2008.

[4] 中村聡史, 山本岳洋, 田中克己. ページ中の感覚情報を利用したウェブ検索支援. 情報処理学会研究会ヒューマンコンピュータインタラクション (2008-HCI123MUS75), pp. 111–116, May 2008.

[5] Michael Wilson, Michael Wilson, Oxon Ox Qx, and Philip Quinlan. Mrc psycholinguistic database: Machine usable dictionary, version 2.00., 1987.