

Web ページの具体度の評価と具体的な Web ページのためのクエリ推薦

田中 伸弥[†] アダム ヤトフト[†] 田中 克己[†]

[†] 京都大学大学院情報学研究科 〒 606-8501 京都府京都市左京区吉田本町

E-mail: [†]{s.tanaka,adam,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし 本稿では Web 文書の具体性の推定方法および、Web 検索エンジンから具体的に記述された Web 文書を効率よく検索するためのクエリ推薦方法を提案する。本稿ではまず単語の具体性の評価手法を提案する。Paivio らにより提案された単語の具象性と心像性に着目し、機械学習の一手法であるサポートベクターマシンを用いた単語の具体性の評価手法を提案する。続いて、単語の具体性を用いた Web 文書の具体度の評価手法を提案する。最後に、具体的な Web 文書集合と抽象的な Web 文書集合から、具体的な Web 文書集合が持つ特徴的な単語を用いた、具体的な Web 文書を効率よく検索できるクエリを推薦する手法を提案する。

キーワード 情報検索，具体性，具象性，心像性

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 [1]。

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 文書のリストを出力する。

既存の Web 検索エンジンではランキングに Web 文書の記述の具体度は考慮されていない。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 文書と、入力クエリの具体的な記述を含む Web 文書が混在している。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 文書を探すのは容易ではない。

既存の Web 検索エンジンに対して、ユーザが具体的な記述を持つ Web 文書を取得したいとき、ユーザは「例」や「方法」などのキーワードを入力クエリに付け加えて検索を行う事がある。これらのキーワードを付け加えることにより、具体的な記述を持つ Web 文書が検索結果に表示されやすいという、経験に基づくためである。どのようなキーワードを付加するかはユーザの経験に基づいており、ユーザの Web 検索能力や Web 検索エンジンにより異なるのが現状である。また、追加するキーワードは入力クエリそのものにも依存している。この様に現在の Web 検索エンジンを用いて具体的な記述を含む Web 文書の検索は、ユーザの経験や Web 検索能力に依存している。本稿では、このようなユーザの経験や Web 検索能力によらない、具体的な記述を持つ Web 文書の発見を目標とする。

本稿ではまず、関連研究を示す。続いて、単語の具体度の評価手法を提案する。単語の具体度を求めるため、Paivio ら [2] により提案された単語の具象性と心像性に着目する。機械学習

の一手法であるサポートベクターマシン (SVM) の線形回帰により単語の具象性および心像性を求め、その後単語の具体度の評価手法を提案する。続いて、単語の具体度を用いた文書の具体度の評価手法を提案する。最後に、具体的な Web ページを効率よく検索できるクエリの推薦手法を提案する。

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。

若木ら [4] は「検索語より具体性があり、検索語から連想するものとして適切であるが、検索語に包含される様々なトピックを網羅する」性質として、Tangibility という指標を提案し、単語の具体性を数値的に測る手法を提案した。

中村ら [5] 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 [2] 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。

3. 具体度の評価

本節では、(1) SVM の線形回帰により単語の具体度を評価する手法、(2) 文書の具体度を求める手法、(3) 具体的な Web ページを検索するためのクエリ推薦手法について述べる。

3.1 単語の具体度の評価

本項では単語の具体度を評価する手法を提案する。単語の具体性を評価するために、Paivio らにより提案された、単語の

具象性 (concreteness) と心像性 (imageability) に着目する [2] . 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. 単語 t の具象性を $\text{Conc}(t)$, 心像性を $\text{Imag}(t)$ とする .

具象性 $\text{Conc}(t)$ と心像性 $\text{Imag}(t)$ を, 単語の具象性と心像性を, 機械学習の一手法であるサポートベクターマシン (SVM) の線形回帰により求める. SVM の線形回帰により求めた単語 t の具象性を $\text{Conc}_{\text{SVM}}(t)$ とし, 単語 t の心像性を $\text{Imag}_{\text{SVM}}(t)$ とする. このとき, 単語 t の具体度 $\text{Conc}_{\text{CSVM}}(t)$ を以下のようにして求める .

$$\text{Conc}_{\text{CSVM}}(t) = \frac{1}{2} (\text{Conc}_{\text{SVM}}(t) + \text{Imag}_{\text{SVM}}(t)) \quad (1)$$

SVM の線形回帰に用いる特徴を, 単語の出現頻度, 関連記述の多様さ, 五感を表す動詞との共起, 意味数, 概念階層での深さ, 下位語数, 印象, および単語長の 8 つのカテゴリから, 21 個の特徴を選出した. 表 1 に SVM の線形回帰に使用した 21 個の特徴を示す .

表 1 SVM の線形回帰に用いた 21 個の特徴

f1	$\text{Popularity}_{\text{web}}(t)$	Web における t の出現頻度
f2	$\text{Popularity}_{\text{image}}(t)$	t に関連した画像数
f3	$\text{Popularity}_{\text{photo}}(t)$	t に関連した写真数
f4	$\frac{\text{Popularity}_{\text{image}}(t)}{\text{Popularity}_{\text{web}}(t)}$	f2 と f1 の比
f5	$\frac{\text{Popularity}_{\text{photo}}(t)}{\text{Popularity}_{\text{web}}(t)}$	f3 と f1 の比
f6	$\#\text{Annotations}(t)$	関連記述の数
f7	$\#\text{Annotations}_{\text{uniq}}(t)$	関連記述の種類数
f8	$\text{SenseVerbs}_{\text{max}}(t)$	五感を表す動詞との共起度の最大値
f9	$\text{SensesVerbs}_{\text{avg}}(t)$	五感を表す動詞との共起度の平均値
f10	$\#\text{Senses}(t)$	意味数
f11	$\text{Depth}_{\text{freq}}(t)$	最も使用頻度の高い意味の概念階層での深さ
f12	$\text{Depth}_{\text{avg}}(t)$	概念階層での深さの平均値
f13	$\#\text{Hyponyms}_{\text{freq}}(t)$	最も使用頻度の高い意味の下位語数
f14	$\#\text{Hyponyms}_{\text{avg}}(t)$	下位語数の平均値
f15	$\text{Positivity}_{\text{freq}}(t)$	最も使用頻度の高い意味の肯定度
f16	$\text{Positivity}_{\text{avg}}(t)$	肯定度の平均値
f17	$\text{Negativity}_{\text{freq}}(t)$	最も使用頻度の高い意味の否定度
f18	$\text{Negativity}_{\text{avg}}(t)$	否定度の平均値
f19	$\text{Objectivity}_{\text{freq}}(t)$	最も使用頻度の高い意味の客観度
f20	$\text{Objectivity}_{\text{avg}}(t)$	客観度の平均値
f21	$\#\text{Characters}(t)$	文字数

3.1.1 出現頻度

単語の心像性は想像のしやすさを表すため, 一般的に広く知られ, 使用されている単語ほど心像性は高くなると考えられる. また, 具象性は人間が五感で感じることができる指標であるので, 単語 t の表す対象を写真に撮ることができるならば, 単語 t の具象性は高いと考えられる. そこで, 単語 t の Web 上での知名度を $\text{Popularity}_{\text{web}}(t)$, 単語 t の画像に対する関連度を $\text{Popularity}_{\text{image}}(t)$, 単語 t の写真に対する関連度を $\text{Popularity}_{\text{photo}}(t)$ とする. これらの特徴を Bing ウェ

ブ検索^(注1)での単語 t の検索結果数 $\text{Freq}_{\text{web}}(t)$, Bing 画像検索での検索結果数 $\text{Freq}_{\text{image}}(t)$, Flickr^(注2)での検索結果数 $\text{Freq}_{\text{photo}}(t)$ を用いて以下のように測定する .

$$\begin{aligned} \text{f1} \quad \text{Popularity}_{\text{web}}(t) &= \log_{10}(\text{Freq}_{\text{web}}(t) + 1) \\ \text{f2} \quad \text{Popularity}_{\text{image}}(t) &= \log_{10}(\text{Freq}_{\text{image}}(t) + 1) \\ \text{f3} \quad \text{Popularity}_{\text{photo}}(t) &= \log_{10}(\text{Freq}_{\text{photo}}(t) + 1) \end{aligned}$$

また, 心像性が高い単語に関しては, 単語 t の Web ページ数に対する, 単語 t に関連した画像の数や, 単語 t に関連した写真の数との比は高いことが考えられる. そのため以下の 2 つの特徴も追加する .

$$\begin{aligned} \text{f4} \quad \frac{\text{Popularity}_{\text{image}}(t)}{\text{Popularity}_{\text{web}}(t)} &= \frac{\log_{10}(\text{Freq}_{\text{image}}(t) + 1)}{\log_{10}(\text{Freq}_{\text{web}}(t) + 1)} \\ \text{f5} \quad \frac{\text{Popularity}_{\text{photo}}(t)}{\text{Popularity}_{\text{web}}(t)} &= \frac{\log_{10}(\text{Freq}_{\text{photo}}(t) + 1)}{\log_{10}(\text{Freq}_{\text{web}}(t) + 1)} \end{aligned}$$

3.1.2 関連記述の多様さ

単語 t が抽象的である場合, 単語 t に関連した記述が多様であると考えられる. 単語 t に関連した記述の多さを $\#\text{Annotations}(t)$ とし, 単語 t に関連した記述の種類数を $\#\text{Annotations}_{\text{uniq}}(t)$ とする .

これら単語 t に関連した記述の多さと種類数を, 単語 t に関連した写真に付けられたタグを用いて測定する .

まず Flickr を単語 t で検索し, 画像集合 $\text{Photos}(t) = \{photo_1, \dots, photo_n\}$ を取得する. ここで, $0 \leq n \leq 500$ とする. そして, $\#\text{Tags}(t)$, $\#\text{Tags}_{\text{uniq}}(t)$ をそれぞれ, $\text{Photos}(t)$ 内の画像に付けられたタグの総数と種類数とする. ただし, $\#\text{Tags}(t)$ と $\#\text{Tags}_{\text{uniq}}(t)$ は, ストップワードとクエリ t 自身はカウントしない. これらの値を基に, 単語 t の関連記述の多様さを以下のようにする .

$$\begin{aligned} \text{f6} \quad \#\text{Annotations}(t) &= \frac{\#\text{Tags}(t)}{n} \\ \text{f7} \quad \#\text{Annotations}_{\text{uniq}}(t) &= \frac{\#\text{Tags}_{\text{uniq}}(t)}{n} \end{aligned}$$

3.1.3 五感を表す動詞との共起

具象性の高い単語 t は, 人間の五感を表す動詞 (see, touch, taste, smell, hear) と共起しやすいと考えられる. 単語 t の視覚を表す動詞との共起頻度を $\text{Cooc}_{\text{see}}(t)$ とし, 単語 t の触覚, 味覚, 嗅覚, 聴覚を表す動詞との共起頻度をそれぞれ $\text{Cooc}_{\text{touch}}(t)$, $\text{Cooc}_{\text{taste}}(t)$, $\text{Cooc}_{\text{smell}}(t)$, $\text{Cooc}_{\text{hearing}}(t)$ とする .

視覚を表す動詞を $\text{Verbs}_{\text{sight}} = \{\text{see, sees, saw, seen}\}$ とする. このとき, 単語 t の視覚を表す動詞との共起頻度 $\text{Cooc}_{\text{sight}}$ を,

$$\text{Cooc}_{\text{sight}}(t) = \sum_{v \in \text{Verbs}_{\text{sight}}} \frac{\text{Freq}_{\text{web}}(“v * t”)}{\text{Freq}_{\text{web}}(v \vee t)}$$

として測定する. 触覚, 味覚, 嗅覚, 聴覚についても同様にして求める. 表 2 に, 五感を表す動詞との共起頻度を求める際に使用した動詞を示す .

(注1): <http://www.bing.com/>

(注2): <http://www.flickr.com>

表 2 五感を表す動詞との共起頻度の測定に使用した動詞

視覚	see, sees, saw, seen
触覚	touch, touches, touched
味覚	taste, tastes, tasted
嗅覚	smell, smells, smelled, smelt
聴覚	hear, hears, heard

これら $\text{Cooc}_{\text{sight}}(t)$, $\text{Cooc}_{\text{touch}}(t)$, $\text{Cooc}_{\text{taste}}(t)$, $\text{Cooc}_{\text{smell}}(t)$, $\text{Cooc}_{\text{hearing}}(t)$ の 5 つの共起頻度のうち、最大値 $\text{Senses}_{\text{max}}(t)$ と平均値 $\text{Senses}_{\text{avg}}(t)$ の 2 つの値を SVM の線形回帰の特徴用いる。すなわち、

$$\text{f8} \quad \# \text{SenseVerbs}_{\text{max}}(t) = \max(\text{Cooc}_{\text{sight}}(t), \text{Cooc}_{\text{touch}}(t), \text{Cooc}_{\text{taste}}(t), \text{Cooc}_{\text{smell}}(t), \text{Cooc}_{\text{hearing}}(t))$$

$$\text{f9} \quad \# \text{SenseVerbs}_{\text{avg}}(t) = \frac{1}{5}(\text{Cooc}_{\text{sight}}(t) + \text{Cooc}_{\text{touch}}(t) + \text{Cooc}_{\text{taste}}(t) + \text{Cooc}_{\text{smell}}(t) + \text{Cooc}_{\text{hearing}}(t))$$

の 2 つの共起頻度を SVM の線形回帰の特徴に用いる。

3.1.4 意味数

同一の単語であっても、複数の意味を持つ場合がある。WordNet^(注3)によれば、名詞 *plant* は植物や工場などの 4 つの異なる意味を持つ。単語 t が多くの意味を意味を持つほど、単語 t はあいまいであり、抽象的であると考えられるため、単語 t の意味数 $\# \text{Senses}(t)$ を SVM の線形回帰の特徴 (f10) に用いる。

3.1.5 概念階層における深さ

具体性の定義に、概念階層での深さをを用いる場合がある [6]。そのため、概念階層における深さを SVM の線形回帰の特徴に加える。

通常単語 t は複数の意味を持ち、それぞれの意味で概念階層における位置が異なる。最も使用頻度の高い意味における概念階層の根からの深さを $\text{Depth}_{\text{freq}}(t)$ とし (f11)、単語 t がとりうる概念階層の根からの深さの平均値を $\text{Depth}_{\text{avg}}(t)$ とする (f12)。これら 2 つの特徴を SVM の線形回帰の特徴に用いる。

3.1.6 下位語数

単語 t の下位語数が多いほどより抽象的であると考えられる。通常単語 t は複数の意味を持ち、それぞれの意味において下位語の数が異なる。最も使用頻度の高い意味における下位語数を $\# \text{Hyponyms}_{\text{freq}}(t)$ とし (f13)、単語 t がとりうる下位語数の平均値を $\# \text{Hyponyms}_{\text{avg}}(t)$ とする (f14)。これら 2 つの特徴を SVM の線形回帰の特徴に用いる。

3.1.7 印象

単語の中には、良い印象や悪い印象といった、主観的な印象をもたらすものがある。SentiWordNet^(注4)によると、*opportunity* は良い印象、*regret* は悪い印象を持つとされる。具体的な単語ほど、客観的な印象を持ち、肯定や否定などの主観的な印象が持ちにくいと考えられる。

通常単語 t は複数の意味を持ち、それぞれの意味において

印象が異なる。最も使用頻度の高い意味における肯定度、否定度、客観度をそれぞれ $\text{Positivity}_{\text{freq}}(t)$, $\text{Negativity}_{\text{freq}}(t)$, $\text{Objectivity}_{\text{freq}}(t)$ とする (f15, f16, f17)。単語 t がとりうる肯定度、否定度、客観度の平均値をそれぞれ $\text{Positivity}_{\text{avg}}(t)$, $\text{Negativity}_{\text{avg}}(t)$, $\text{Objectivity}_{\text{avg}}(t)$ とする (f18, f19, f20)。これら 6 つの特徴を SVM の線形回帰の特徴に用いる。

3.1.8 単語長

抽象的な単語は *-tion* や *-bility* などの接尾語が付きやすく、これら接尾語により、抽象語ほど単語 t を構成している文字数が多くなると考えられる。単語 t を構成している文字数を $\# \text{Characters}(t)$ とし (f21)、SVM の線形回帰の特徴に用いる。

3.2 文書の具体度の評価

本項では、文書内の単語の具体度を用いて、文書の具体度の評価手法を述べる。本稿では文書の具体度を評価するにあたり、2 つの仮説を立てた。1 つ目が、1 つの文書の中にも、具体的な段落と、抽象的な段落とが混在しているということ。2 つ目が、具体的な段落は主に具体的な単語からなり、抽象的な段落は主に抽象的な単語からなることである。これらの仮説のもと、文書中の各段落に対して段落の具体度を計算し、段落の具体度を用いて文書の具体度として求める。

文書 D を段落の順列 $D = (P_1, P_2, \dots, P_N)$ とする。ここで、 P_i は文書 D 中の段落を表す。各段落 P_i は段落中の単語の順列 $P_i = (t_1, t_2, \dots, t_M)$ とする。文書 D 中のそれぞれの段落 P に対して、段落の具体度 $\text{Conc}_{\text{par}}(P)$ を次の式で求める。

$$\text{Conc}_{\text{par}}(P) = \frac{1}{|P|} \sum_{t \in P} \text{Conc}(t) \quad (2)$$

式 2 では段落の長さを考慮に入れていない。極端に短い段落は、広告などのノイズである可能性が高い。段落の長さを考慮に入れた、段落の具体度を以下に示す。

$$\text{Conc}'_{\text{par}}(P) = (1 - \alpha^{-|P|}) \text{Conc}_{\text{par}}(P) \quad (3)$$

ここで α は 0 もしくは 1 以上の値をもつパラメータである。 α が 0 のとき $\text{Conc}'_{\text{par}}(P)$ と $\text{Conc}_{\text{par}}(P)$ は等しくなる。

最後に、文書 D の具体度 $\text{Conc}_{\text{doc}}(D)$ を次の式で求める。

$$\text{Conc}_{\text{doc}}(D) = \max(\text{Conc}'_{\text{par}}(P_1), \dots, \text{Conc}'_{\text{par}}(P_N))$$

文書の具体度の評価方法として、単純に、文書中の単語の具体度の平均値により求めることが考えられる。しかし、文書の一部に、抽象的な概念の例やインスタンスが書かれている、という場合が考えられるため、上述のように文書中の段落の具体度の最大値を用いて、文書の具体度の評価を行った。

3.3 具体的な Web ページの検索のためのクエリ推薦

本項では、具体的なコンテンツを持つ Web ページを検索するためのクエリ推薦手法を述べる。まず、3.3.1 目で問題を定義し、3.3.2 目でクエリ推薦手法を述べる。

3.3.1 問題定義

入力クエリを q とし、クエリ q に対する検索結果を $\text{Results}(q) = (D_1, D_2, \dots, D_K)$ とする。ここで D_i は Web

(注3): <http://wordnet.princeton.edu/>

(注4): <http://sentiwordnet.isti.cnr.it>

ページを表す．

クエリ推薦の目的は，入力クエリ q の検索結果 $\text{Results}(q)$ よりも，検索結果 $\text{Results}(q')$ に具体的な Web ページを多く含むクエリ q' を求めることである．

多くの場合， q' の検索結果 $\text{Results}(q')$ 内の Web ページ p のコンテンツは，クエリ q に関連した抽象的なトピックと instance-of の関係により表現される．例えば，法律に対しては判例，パーキンソン病に対してはパーキンソン病の症状やパーキンソン病と闘病生活にある人々の体験談などは instance-of の関係により表現される．

本稿では推薦するクエリに，入力クエリにキーワードを追加した拡張クエリを用いる．クエリ q に対して推薦されたキーワード集合を $\text{Keywords}(q)$ とする．クエリ q の検索結果のトピックを $\text{Topic}_{\text{res}}(t)$ とする．このとき，問題は以下の条件を満たすキーワード集合 $\text{Keywords}(q)$ を求めることである．

$$\begin{cases} \text{Topic}_{\text{res}}(q) \simeq \text{Topic}_{\text{res}}(q+k) \\ \text{Conc}_{\text{res}}(q) < \text{Conc}_{\text{res}}(q+k) \end{cases} \quad \text{for } k \in \text{Keywords}(q)$$

ここで， $\text{Conc}_{\text{res}}(q)$ は以下のとおりである．

$$\text{Conc}_{\text{res}}(q) = \frac{1}{|\text{Results}(q)|} \sum_{D \in \text{Results}(q)} \text{Conc}_{\text{doc}}(D)$$

3.3.2 手 法

まず，入力クエリ q に対して検索結果上位 N 件の Web ページを取得する．取得した Web ページ集合を $\text{Results}(q)$ とする．次に， $\text{Results}(q)$ 内のそれぞれの文書 D に対して，具体度 $\text{Conc}_{\text{doc}}(D)$ を計算する．

次に，推薦するキーワードの候補集合 $\text{Terms}(q)$ を $\text{Results}(q)$ から得る．候補集合 $\text{Terms}(q)$ 内の単語 t に対して， $\text{Docs}(q \wedge t)$ を， $\text{Results}(q)$ 中の単語 t を含む文書の集合とする．同様に， $\text{Docs}(q \vee t)$ を， $\text{Results}(q)$ 中の単語 t を含まない文書の集合とする．

$\text{Terms}(q)$ 内の異なる単語のペア t_i と t_j に対して， $\text{Docs}(q \wedge t_i)$ と $\text{Docs}(q \wedge t_j)$ が等しい場合は， t_i と t_j のうち，入力クエリ q と関連度の高い単語のみを使用する．クエリ q と単語 t との関連度を $\text{Relation}(q, t)$ を次の式で求める．

$$\text{Relation}(q, t) = \frac{\text{Freq}_{\text{web}}(q \wedge t)}{\text{Freq}_{\text{web}}(q \vee t)}$$

また， $|\text{Docs}(q \wedge t)|$ が閾値 θ よりも小さい単語 t に対しても， $\text{Terms}(q)$ から取り除く．

次に，単語 t の具体性への貢献度 $\text{Contrib}(q, t)$ を次の式で求める．

$$\begin{aligned} \text{Contrib}(q, t) = & \frac{1}{|\text{Docs}(q \wedge t)|} \sum_{D \in \text{Docs}(q \wedge t)} \text{Conc}_{\text{doc}}(D) \\ & - \frac{1}{|\text{Docs}(q \wedge \neg t)|} \sum_{D \in \text{Docs}(q \wedge \neg t)} \text{Conc}_{\text{doc}}(D) \end{aligned}$$

また，文書集合 $\text{Docs}(q \wedge t)$ と $\text{Results}(q)$ との $tf-idf$ ベクトルのセントロイドによるコサイン類似度を $\text{TopicSim}(q, t)$ と

する．

最後に，候補単語の集合 $\text{Terms}(q)$ 内の単語 t の推薦度 $\text{Rec}(q, t)$ を次の式で求める．

$$\text{Rec}(q, t) = \text{Contrib}(q, t) \times \text{TopicSim}(q, t) \quad (4)$$

候補単語の集合 $\text{Terms}(q)$ 内の単語のうち，推薦度の高い単語を入力クエリ q に追加するキーワードとして推薦する．

4. 評価と分析

本節では，提案手法の実験の設定および実験結果を示す．本実験は 2011 年 10 月から 2012 年 1 月にかけて行われた．

4.1 データセット

実験には主なデータセットとして，Medical Research Council Psycholinguistic Database(MRCDB) version 2.08 を使用し，WordNet 3.0，Bing，Flickr，および SentiWordNet 3.0 を SVM の線形回帰用いる各特徴を求めるために使用した．

MRCDB ではユーザ実験による単語の具象性と心像性の値が記録されている．MRCDB は主として心理言語学者のために作成され，150,837 個の単語に対して 26 個の異なる単語の属性値が記録されている．ただし，すべての単語に対して 26 個すべての属性値が付与されていない．MRCDB に記録されている具象性と心像性は，Paivio [2]，Toglia，Batting [7]，および Gilhooly，Logie [8] の 3 つの研究による測定値に基づいている．これら 3 つの研究による測定値には高い相関があり，平均値と標準偏差による調整を行ったのち，1 つにまとめられた．MRCDB では，8,288 個の単語に対して具象性が，9,240 個の単語に対して心像性の値が付与されている．ユーザ実験による具象性と心像性の値は [1, 7] の区間の値を取るが，MRCDB では単語の具象性と心像性は [100, 700] の区間の値をとる．

具象性の評価は，1. 非常に抽象的 (単語の対象を経験・知覚する事ができない)， $\dots$ ，7. 非常に具体的 (単語の対象を容易に経験・知覚することができる)，という 7 段階のユーザ実験により求められる．同様に心像性の評価は，1. イメージは無い (単語の対象を想像する事が不可能，もしくは想像に非常に時間がかかる)， $\dots$ ，7. 見えているように鮮明 (単語の対象を容易に素早く想像できる)，という 7 段階のユーザ実験により求められる．MRCDB における具象性と心像性の値は 3 つの研究での測定値に基づくが，3 つの研究のうち 1 つ Paivio [2] では，単語の具象性は 28 人 (男性 12 人)，心像性は 30 人 (男性 15 人) の評価者による評価値である．

MRCDB での単語 t の具象性を $\text{Conc}_{\text{MRC}}(t)$ ，心像性を $\text{Imag}_{\text{Imag}}(t)$ とする．これら $\text{Conc}_{\text{MRC}}(t)$ と $\text{Imag}_{\text{SVM}}(t)$ は [100, 700] の区間の値を取るが，最大値 700 と最小値 100 を用いて，[0, 1] の区間にスケールした値を使用する．

具象性と心像性の組み合わせた単語の具体度 $\text{Conc}_{\text{Cans}}(t)$ を次の式により求める．

$$\text{Conc}_{\text{Cans}}(t) = \frac{1}{2} \{ \text{Conc}_{\text{MRC}}(t) + \text{Imag}_{\text{MRC}}(t) \}.$$

本稿での実験では，MRCDB で具象性と心像性の値を共に持ち，なおかつ WordNet に名詞として登録されている単語に焦

点をあてる．3,455 個の名詞がこれらの条件を満たす．表 3 に、それら 3,455 個の名詞に対する MRCDB における具象性と心像性の最小値，最大値，平均値，および標準偏差を示す．表 4

表 3 MRCDB の 3,455 個の名詞の統計値

尺度	min	max	Avg.	S.D.
Conc (t)	0.097	0.950	0.578	0.195
Imag (t)	0.048	0.945	0.620	0.175
Conc _{ans} (t)	0.152	0.923	0.609	0.177

に具象性，心像性，およびそれらを組み合わせた値の上位 10 個の単語と下位 10 個の単語を示す．MRCDB に登録されている

表 4 具象性と心像性の上位下位 10 位までの単語

具象性上位 10 位までの単語: 単語 t (Conc (t))
milk(0.950), tomato(0.937), grasshopper(0.933), ape(0.923), pineapple(0.922), cucumber(0.922), skunk(0.913), cement(0.910), steak(0.910), beet(0.908)
具象性下位 10 位までの単語: 単語 t (Conc (t))
as(0.097), are(0.158), unreality(0.172), infinity(0.173), while(0.173), incongruity(0.178), immortality(0.182), tun(0.185), sagacity(0.185), indolence(0.185)
心像性上位 10 位までの単語: 単語 t (Imag (t))
beach(0.945), jeep(0.932), telephone(0.925), sunshine(0.925), skunk(0.920), cranberry(0.915), steak(0.912), father(0.910), cigarette(0.908), banana(0.907)
心像性下位 10 位までの単語: 単語 t (Imag (t))
plenipotentiary(0.048), tush(0.072), astrolabe(0.083), convocation(0.085), equanimity(0.088), wherry(0.093), philology(0.103), provender(0.103), incursion(0.107), yucca(0.112)
組み合わせた具体度の上位 10 位までの単語: 単語 t (Conc _{ans} (t) (Conc (t), Imag (t)))
milk(0.92(0.95, 0.90)), skunk(0.92(0.91, 0.92)), steak(0.91(0.91, 0.91)), grasshopper(0.91(0.93, 0.88)), jeep(0.90(0.87, 0.93)), beach(0.90(0.85, 0.95)), banana(0.90(0.89, 0.91)), cucumber(0.90(0.92, 0.87)), telephone(0.90(0.87, 0.93)), tomato(0.89(0.94, 0.85))
組み合わせた具体度の下位 10 位までの単語: 単語 t (Conc _{ans} (t) (Conc (t), Imag (t)))
as(0.15(0.10, 0.21)), equanimity(0.15(0.22, 0.09)), convocation(0.17(0.26, 0.09)), incongruity(0.18(0.18, 0.18)), impunity(0.19(0.24, 0.14)), tush(0.19(0.31, 0.07)), monad(0.19(0.25, 0.14)), philology(0.20(0.29, 0.10)), nabob(0.20(0.21, 0.19)), indolence(0.20(0.19, 0.21))

3,455 個の名詞の具象性と心像性のピアソンの相関係数は 0.826 であり，具象性と心像性には強い正の相関があることがわかる．

MRCDB 内での具象性と心像性の値を，SVM の線形回帰の正解データとして用いた．

SVM の線形回帰のために，各特徴の値は，特徴内での最大値と最小値を用いて， $[0, 1]$ の区間にスケーリングしたものをを用いた．

2 つの値の順列間の相関を評価するために，ピアソンの相関

係数，ケンドールのタウ，平均二乗平方根誤差 (RMSE) の 3 つの尺度を用いた．ピアソンの相関係数とケンドールのタウは $[-1, 1]$ の区間の値を取る．それらの値が大きいもしくは小さいとき，入力 of 2 つの値の順列間には相関があることを示唆し，値が 0 のときは相関が無いことを示唆する．平均二乗平方根誤差が小さい時，入力 of 2 つの値の順列間の差は小さいことを示唆する．

4.1.1 結果

SVM^{light} [9] を標準的なパラメータで用いて，単語の具象性 Conc_{SVM} (t) と心像性 Imag_{SVM} (t) を推定した．そして，式 1 で表した単語の具体度を，SVM の線形回帰と 5-分割交差検定を用いて推定した．

ユーザ実験による具象性の値 Conc (t) と SVM による推定値 Conc_{SVM} (t) とのピアソンの相関係数，ケンドールのタウ，および平均二乗平方根誤差はそれぞれ，0.671，0.492，0.145 であった．ユーザ実験による心像性の値 Imag (t) と SVM による推定値 Imag_{SVM} (t) においてはそれぞれ，0.675，0.502，0.129 であった．心像性のほうが具象性よりも推定精度がよいことがわかる．MRCDB 内の具象性と心像性の値による具体度 Conc_{ans} (t) と式 1 で求めた Conc_{CSVM} (t) については，それぞれ，0.688，0.508，0.128 という結果を得た．表 5 に Conc_{SVM} (t)，Imag_{SVM} (t)，Conc_{CSVM} (t) の統計値を示す．SVM の線形回帰により求めた

表 5 SVM の線形回帰の結果の統計値

尺度	min	max	Avg.	S.D.
Conc _{SVM} (t)	-0.016	1.010	0.601	0.132
Imag _{SVM} (t)	0.105	1.017	0.623	0.116
Conc _{CSVM} (t)	0.085	1.001	0.612	0.122

具象性 Conc_{SVM} (t) は，ユーザ実験による具象性 Conc (t) よりも標準偏差が小さい事がわかった．心像性にも同様の傾向がみられ，これらは平滑化による影響と考えられる．

表 6 に具象性，心像性，およびそれらを組み合わせた値の上位 10 個の単語と下位 10 個の単語を示す．また，SVM の線形回帰により推定された具象性 Conc_{SVM} (t) と心像性 Imag_{SVM} (t) のピアソンの相関係数，ケンドールのタウ，平均二乗平方根誤差はそれぞれ，0.919，0.764，0.057 であった．推定された具象性と心像性には，ユーザ実験による値よりも強い正の相関があることという結果を得た．

4.1.2 特徴の重要度の評価

本目では，SVM の線形回帰に用いた特徴の重要度を，5-分割交差検定と，ピアソンの相関係数，ケンドールのタウ，平均二乗平方根誤差 (RMSE) の 3 つの尺度を用いて評価した．

特徴の重要度の評価は，特徴集合 *Features* から，各評価尺度で最も重要でない特徴 f を見つけ出し *Features* から取り除く，という処理を段階的に行い評価する．

表 7 に，ピアソンの相関係数，ケンドールのタウ，平均二乗平方根誤差の平均値の推移を示す．ステップ 0 の値は 21 個すべての特徴を用いた場合の値である．表 7 の各列は，取り除かれた特徴と，線形回帰により推定された値と正解の値との評価値である．単語のもつ客観度の平均値である特徴 f_{20} が具象性

表 7 ピアソン相関係数，ケンドールのタウ，平均二乗平方根誤差 (RMSE) の推移

step	ConcSVM (t)			ImagSVM (t)		
	ピアソン	ケンドール	RMSE	ピアソン	ケンドール	RMSE
0	– 0.668	– 0.489	– 0.145	– 0.671	– 0.499	– 0.130
1	f5 0.669	f14 0.490	f5 0.145	f5 0.673	f6 0.499	f5 0.129
2	f14 0.669	f17 0.490	f14 0.145	f1 0.673	f20 0.499	f1 0.129
3	f16 0.669	f8 0.490	f16 0.145	f9 0.673	f8 0.499	f19 0.129
4	f19 0.669	f9 0.490	f19 0.145	f14 0.673	f1 0.500	f17 0.129
5	f8 0.669	f2 0.490	f8 0.145	f13 0.673	f5 0.500	f8 0.129
6	f9 0.670	f18 0.490	f9 0.145	f20 0.673	f9 0.500	f13 0.129
7	f2 0.670	f15 0.490	f2 0.145	f18 0.674	f17 0.500	f14 0.129
8	f6 0.669	f5 0.490	f18 0.145	f8 0.674	f13 0.500	f6 0.129
9	f18 0.669	f6 0.489	f17 0.145	f19 0.674	f15 0.500	f16 0.129
10	f17 0.669	f19 0.488	f6 0.145	f6 0.674	f14 0.500	f9 0.129
11	f13 0.667	f13 0.487	f13 0.146	f15 0.674	f19 0.500	f15 0.129
12	f10 0.665	f10 0.485	f10 0.146	f21 0.673	f12 0.498	f21 0.129
13	f15 0.663	f16 0.483	f15 0.146	f17 0.671	f18 0.496	f18 0.130
14	f12 0.660	f12 0.479	f12 0.147	f12 0.669	f4 0.492	f12 0.130
15	f7 0.652	f7 0.470	f7 0.148	f10 0.666	f21 0.489	f4 0.130
16	f21 0.635	f21 0.459	f21 0.151	f4 0.664	f10 0.486	f10 0.131
17	f4 0.619	f4 0.443	f4 0.153	f16 0.653	f16 0.474	f20 0.132
18	f1 0.592	f1 0.422	f1 0.157	f7 0.631	f7 0.458	f7 0.136
19	f11 0.532	f11 0.359	f11 0.166	f2 0.608	f11 0.431	f2 0.139
20	f3 0.415	f3 0.287	f3 0.196	f11 0.344	f2 0.230	f11 0.448
21	f20 –	f20 –	f20 –	f3 –	f3 –	f3 –

の推定に最も重要であり，関連写真数である特徴 **f3** が心像性を推定するのに最も重要であることがわかる．特徴 **f3** と，最も使用頻度の高い意味における概念階層での深さである特徴 **f11** が，具象性と心像性，両属性値を推定するのに重要であることがわかる．一方で，感覚動詞との共起のしやすさである特徴 **f9** や下位語数 **f14** などは線形回帰において重要ではないことがわかる．

4.2 クエリ推薦

本項では，提案手法の具体的な Web ページを検索するためのクエリ推薦手法の結果を示す．

実験には，*antitrust law*，*democracy*，*environmental issues*，*financial crisis*，*information security*，*leprosy*，*parkinson's disease*，*self evaluation*，*stealth marketing* および *subprime lending* の 10 個のクエリを用いた．これらのクエリは [10] と今日 Web 上で話題性のあるクエリから選出した．

まず，それぞれのクエリに対して，提案手法の推薦度上位 10 個のキーワードを取得し，初期クエリにそれら推薦されたキーワードを追加した拡張クエリにより Bing で検索し，検索結果を取得した．初期クエリと 10 個の拡張クエリそれぞれに対する，検索結果上位 10 件の Web ページがクエリに対して具体的であるかどうかを，二値で評価した．本実験では合計 1100 の Web ページの具体度を評価した．ページが存在しないなどのエラーページは評価から除外した．

実験の設定として，式 3 の短すぎる段落の影響度を小さくするためのパラメータ α を $e \approx 2.72$ とした．また，取得する Web ページ数 N は 200 とし，出現頻度の低い単語を除外する

ための閾値 θ は $0.1N = 20$ とした．実験を単純にするために，文書の具体度の計算には，WordNet に登録されている名詞のみを用いた．WordNet 3.0 に登録されている 117,798 個のすべての名詞に対して単語の具体度をあらかじめ準備した．

拡張されたクエリによる検索結果上位 10 件のうち具体的な Web ページの数の平均値は 5.1 であった．拡張なしのクエリによる検索結果上位 10 件のうち具体的な Web ページの数の平均値は 3.4 であり，推薦されたキーワードを追加したクエリを用いて検索を行った方が，具体的な Web ページを多く検索できる事がわかった．より厳しい評価として，拡張なしクエリによるトピックと，大きく異なるトピックをもつ Web ページを抽象的な Web ページとして評価した場合，拡張されたクエリによる検索結果上位 10 件のうち具体的な Web ページの数の平均値は 4.5 であった．

表 8 と表 9 に，*parkinson's disease* と *environmental issues* に関する推薦されたキーワードを示す．表 8 の上位に，*michael* と *fox* というキーワードが現れている．これらは，パーキンソン病と診断されたアメリカの俳優である Michael. J. Fox 氏を意味している．クエリ *parkinson's disease* + *michael* で検索することで，彼のパーキンソン病に関するインタビューや活動などの Web ページが取得できる．また，推薦されたキーワードには，*leg* や *limb* などの類義語のペアが見つかった．類義語のペアをどちらか一方に集約するなどの手法により，提案手法の改良が可能であると考えられる．キーワードの推薦度の計算においては，キーワードを追加することによりトピックが変化しないよう，キーワードを含む文書集合と，初期クエリによる

表 8 *Parkinson's disease* に対して推薦されたキーワード

ランク	単語 t	Rec (t)	Docs	Contrib	TopicSim
1	michael	0.0369	35	0.041	0.912
2	fox	0.0325	30	0.036	0.901
3	challenge	0.0313	21	0.039	0.811
4	friend	0.0309	36	0.034	0.916
5	sense	0.0286	22	0.034	0.829
6	face	0.0283	21	0.036	0.782
7	neurologist	0.0268	27	0.029	0.912
8	hand	0.0267	43	0.029	0.936
9	american	0.0264	37	0.028	0.933
10	leg	0.0264	30	0.028	0.926

ランク	単語 t	Rec (t)	Docs	Contrib	TopicSim
100	nerve	0.0113	37	0.013	0.891
101	surgery	0.0113	35	0.012	0.934
102	progression	0.0113	43	0.012	0.942
103	trial	0.0112	35	0.012	0.936
104	link	0.0112	31	0.013	0.834
105	heart	0.0111	35	0.013	0.871
106	sign	0.0111	45	0.012	0.889
107	product	0.0111	24	0.012	0.914
108	answer	0.0105	22	0.014	0.750
109	term	0.0105	31	0.012	0.891

表 9 *environmental issues* に対して推薦されたキーワード

ランク	単語 t	Rec (t)	Docs	Contrib	TopicSim
1	science	0.0310	46	0.033	0.931
2	day	0.0276	33	0.031	0.896
3	book	0.0275	22	0.032	0.865
4	water	0.0274	87	0.028	0.968
5	health	0.0253	74	0.026	0.980
6	green	0.0246	34	0.039	0.632
7	number	0.0245	36	0.026	0.939
8	sustainability	0.0236	24	0.027	0.864
9	department	0.0229	33	0.026	0.887
10	september	0.0226	20	0.037	0.606

ランク	単語 t	Rec (t)	Docs	Contrib	TopicSim
100	community	0.0124	43	0.019	0.640
101	member	0.0123	22	0.022	0.559
102	state	0.0123	64	0.018	0.675
103	need	0.0122	37	0.019	0.655
104	support	0.0122	28	0.019	0.628
105	city	0.0121	20	0.022	0.559
106	public	0.0121	29	0.020	0.614
107	effort	0.0120	39	0.018	0.654
108	specie	0.0120	38	0.019	0.628
109	importance	0.0119	22	0.020	0.606

文書集合との類似度を用いている．しかし多くの場合，クエリによる検索結果にはいくつかのトピックが混在している．例えばパーキンソン病の場合，パーキンソン病の定義，原因，症状などのトピックが含まれる．トピックの変化を防ぐために，初期クエリによる検索結果をクラスタリングし，それぞれのクラスタに対して，本稿でのクエリ推薦手法を用いることで，クエリ推薦手法を改良する事ができると考えられる．

5. 議論

5.1 単語の共起と具体性

本項では，単語の共起を用いた単語の具体性の評価の可能性を述べる．

我々は，抽象的な名詞は具体的な名詞よりも抽象的な名詞と共起しやすく，具体的な名詞は抽象的な名詞よりも具体的な名詞と共起しやすい傾向があると仮説を立てた．

この仮説を調査するため，英語版 Wikipedia^(注5) のすべての記事に対して，同一の文での単語の共起と，単語の具象性と心像性を調査した．

$\#Sentences(t)$ を単語 t を含む文数とする．任意の 2 単語 t と u の Jaccard 係数を以下のようにして求める．

$$Jaccard(t, u) = \frac{\#Sentences(t \wedge u)}{\#Sentences(t \vee u)}$$

単語を具象性を順で並べ，それぞれ上位下位 $x\%$ を具象性の高い単語集合 $Terms_{high}$ ，具象性の低い単語集合 $Terms_{low}$ とする．心像性においても同様とする．2 つの単語集合 $Terms_i$ と $Terms_j$ の共起度を，それぞれの単語集合に属する単語の Jaccard 係数の平均値として求める．図 1 と図 2 に，2 つの単語集合間における共起度の結果を示す．具象性および心像性の高い単語は，具象性および心像性の低い単語よりも，具象性および心像性の高い単語を共起しやすいことを示唆している．単語の共起情報を SVM の特徴に追加する事や，単語の共起情報から単語の具体度の推定が可能であると考えられる．

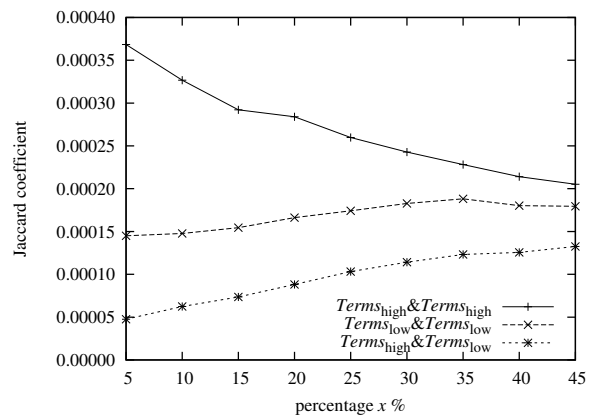


図 1 具象性に関する Jaccard 係数の平均値

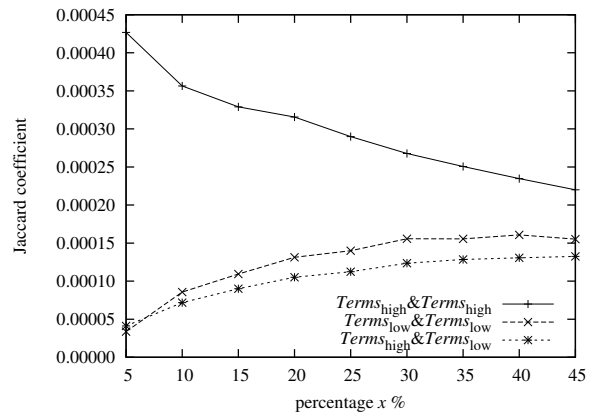


図 2 心像性に関する Jaccard 係数の平均値

5.2 抽象的な Web ページ検索のためのクエリ推薦

本稿では，具体的なコンテンツを持つ Web ページを検索するためのクエリ推薦手法を提案した．同様の手法を用いることで，抽象的なコンテンツを持つ Web ページを検索するためのクエリを推薦する事ができる．具体的な文書は文書長が長すぎ，読むことや覚えることが困難である場合がある．また，検索したい情報を得るのに多くの時間を費やしたくないユーザも存在

(注5): <http://dumps.wikimedia.org/enwiki/20110722/enwiki-20110722-pages-articles.xml.bz2>

表 6 SVM の線形回帰による具象性と心像性の上位下位 10 までの単語

具象性上位 10 位までの単語: 単語 t ($\text{Conc}_{\text{SVM}}(t)$ ($\text{Conc}(t)$))
cattle(1.01(0.83)), portrait(0.99(0.78)), deer(0.94(0.89)), cow(0.94(0.87)), shrub(0.93(0.81)), tree(0.93(0.84)), robin(0.92(0.90)), boat(0.92(0.90)), moose(0.92(0.86)), feline(0.91(0.64))
具象性下位 10 位までの単語: 単語 t ($\text{Conc}_{\text{SVM}}(t)$ ($\text{Conc}(t)$))
competence(-0.02(0.26)), exactitude(0.02(0.25)), ingratitude(0.03(0.25)), comradeship(0.10(0.36)), insufficiency(0.07(0.26)), importance(0.12(0.24)), integrity(0.13(0.25)), infallibility(0.13(0.21)), virtue(0.14(0.24)), misconception(0.14(0.25))
心像性上位 10 位までの単語: 単語 t ($\text{Imag}_{\text{SVM}}(t)$ ($\text{Imag}(t)$))
portrait(1.02(0.78)), cattle(0.94(0.87)), boat(0.92(0.89)), sunlight(0.91(0.91)), cow(0.90(0.89)), grass(0.90(0.84)), dusk(0.90(0.79)), deer(0.90(0.87)), sunset(0.90(0.89)), mare(0.89(0.72))
心像性下位 10 位までの単語: 単語 t ($\text{Imag}_{\text{SVM}}(t)$ ($\text{Imag}(t)$))
besieger(0.11(0.39)), exactitude(0.15(0.33)), disparagement(0.18(0.16)), ingratitude(0.19(0.39)), increment(0.20(0.36)), impotency(0.20(0.51)), competence(0.20(0.40)), profiteer(0.21(0.33)), interposition(0.21(0.17)), loquacity(0.21(0.37))
組み合わせた具体度の上位 10 位までの単語: 単語 t ($\text{Conc}_{\text{CSVM}}(t)$ ($\text{Conc}_{\text{SVM}}(t)$, $\text{Imag}_{\text{SVM}}(t)$))
portrait(1.00(0.99, 1.02)), cattle(0.97(1.01, 0.94)), cow(0.92(0.94, 0.90)), boat(0.92(0.92, 0.92)), deer(0.92(0.94, 0.90)), robin(0.90(0.92, 0.89)), tree(0.90(0.93, 0.88)), grass(0.90(0.90, 0.90)), dusk(0.90(0.90, 0.90)), sunlight(0.90(0.89, 0.91))
組み合わせた具体度の下位 10 位までの単語: 単語 t ($\text{Conc}_{\text{CSVM}}(t)$ ($\text{Conc}_{\text{SVM}}(t)$, $\text{Imag}_{\text{SVM}}(t)$))
exactitude(0.09(0.02, 0.15)), competence(0.09(-0.02, 0.20)), ingratitude(0.11(0.03, 0.19)), insufficiency(0.15(0.07, 0.22)), besieger(0.16(0.21, 0.11)), infallibility(0.18(0.13, 0.22)), impotency(0.19(0.18, 0.20)), condescension(0.20(0.15, 0.24)), loquacity(0.20(0.19, 0.21)), comradeship(0.20(0.10, 0.30))

する．そのような場合，数行で書かれた抽象的なコンテンツが有用であると考えられる．また，具体的な文書は，ユーザがその文書のトピックを理解するのに有用であるが，具体的な文書は，抽象的な文書がもっている情報を失っている場合がある．例えば，パーキンソン病の症状に関する文書について考えると，抽象的な文書には，*motor symptoms* と書かれている．一方で，具体的な文書には，*loss of balance* や *shaking in limbs* などが書かれている．多くの場合，具体的な事例から抽象的な概念や法則を推定するのは困難である．パーキンソン病の症状の例の場合，*motor symptoms* は，*loss of balance* や *shaking in limbs* を包含している．したがって，具体的なコンテンツを持つ Web ページ検索するためのクエリ推薦と同様に，抽象的なコンテンツを持つ Web ページ検索するためのクエリ推薦も有用であると考えられる．

6. ま と め

具体的なコンテンツは，抽象的な概念や，複雑な概念などを理解させるために有用である．しかし，具体的なコンテンツの検索を支援する研究は少ない．本稿では，単語および文書の具体度を，単語の具象性と心像性により評価した．単語の具象性と心像性を推定するにあたり，(1) 出現頻度，(2) 関連記述の多様さ，(3) 感覚動詞との共起のしやすさ，(4) 意味数，(5) 概念階層での深さ，(6) 下位語数，(7) 単語の印象，および (8) 文字数の 8 個のカテゴリから合計 21 個の特徴を選択し，SVM の線形回帰により単語の具象性と心像性を求めた．ユーザによる評価値と，SVM の線形回帰により求めた値とのピアソンの相関係数は，具象性については 0.671，心像性については 0.675 という結果を得た．

本稿の 2 つ目の貢献として，具体的なコンテンツをもつ Web ページを検索するためのクエリ推薦手法を提案した．

本稿で提案した具体性の評価方法は，Web 検索のみならず，文書からの情報抽出や，文書の要約などの他の分野の研究に対しても利用できると考えられる．今後の展望として，文書の具体度と文書の関連度により焦点を当てる予定である．

謝辞 本研究の一部は，グローバル COE 拠点形成プログラム「知識循環社会のための情報学教育研究拠点」(拠点リーダー：田中克己) によるものです．ここに記して謝意を表します．

文 献

- [1] John Friedlander. Abstract, concrete, general, and specific terms. <http://grammar.ccc.commnet.edu/grammar/composition/abstract.htm>.
- [2] ALLAN PAIVIO, JOHN C. YUILLE, and STEPHEN A. MADIGAN. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, Vol. 76, No. 1, Part 2, pp. 1–25, 1968.
- [3] Makoto KATO, Hiroaki OHSHIMA, Satoshi OYAMA, and Katsumi TANAKA. "likely" image search: Web image search using term sets representing typical features extracted from social tagging information. *Data Engineering Workshop 2008(DEWS2008)*, March 2008.
- [4] Hiromi WAKAKI, Tomonari MASADA, Atsuhiko TAKASU, and Jun ADACHI. Exhaustive topic detection and query expansion support based on substance-oriented term clustering. *Data Engineering Workshop 2006(DEWS2006)*, 2006.
- [5] Satoshi NAKAMURA, Takehiro YAMAMOTO, and Katsumi TANAKA. Enhancing web search by using sense on the web. *IPSJ SIG Technical Report*, 2008.
- [6] FaaßT. and Kaczmirek L. Psycholinguistic determinants of question difficulty: A web experiment. In *7th International Conference on Social Science Methodology*, 2008.
- [7] Handbook of semantic word norms. *John Wiley & Sons, Inc., One Wiley Drive, Somerset, New Jersey 18873*, p. 152, 1978.
- [8] K J Gilhooly and R H Logie. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods*, Vol. 12, No. 4, pp. 395–427, 1980.
- [9] Thorsten Joachims. *Making large-scale support vector machine learning practical*, pp. 169–184. MIT Press, 1999.
- [10] Easiest-first search: towards comprehension-based web search. pp. 2057–2060, 2009.