C-13

単語心像性を用いた質問回答文の因子得点の推定精度の向上

Improving Estimation of Factor Scores of Answer Statement by Using Word Imageability

横山 友也†　宝珍 輝尚†　野宮 浩揮†　佐藤 哲司‡
Yuya Yokoyama　Teruhisa Hochin　Hiroki Nomiya　Tetsuji Satoh

1．はじめに

インターネット上において，質問回答サイトの利用者が近年急増している．質問回答サイトとは，インターネット上でユーザー同士が互いに質問と回答を投稿しあうコミュニティの一種であり，様々な悩み事・相談事を解決する場であると同時に，膨大な知識が蓄積されたデータベースとして活用されるようにもなってきている．あるユーザーが質問を投稿すると，他のユーザーがその質問に対して回答を投稿する．質問者は，質問文に対して最も適切と判断した回答文を「ベストアンサー」に選定し，その回答を行った回答者に謝礼として手持ちのポイントを贈与する．ここで，「ベストアンサー」とは，質問文に対する満足度が最も高いと質問者が主観的に判断した回答文である．

質問回答サイトの参加者が増え，また，投稿される質問数が膨大になると，回答者が自身の専門性や興味に合った適切な質問文を探し出すことが困難になるという問題が顕在化してくる．あるユーザーが質問文を投稿しても，その質問文が必ずしも適切な回答者の目に留まり，回答を得られるわけではないという問題である．また，適切な回答者に巡り会えないミスマッチから，質問者にも不利益も生じる．つまり，質問回答サイトの課題は，日々投稿され続けている幾多の質問と，様々な興味・関心や専門性を有する回答者とを適切にマッチングすることであるが，質問者や回答者の努力に任せているのが現状である．そこで，ある質問文に適切な回答ができるユーザーをその質問文に引き合わせるための方法が研究されている．

これまでの研究において，筆者らは，質問者に適切な回答者を引き合わせるために，質問者と回答者の相性を判断する手段として質問者と回答者の文章の印象評価を行ってきた．50 個の印象語を使用して，Yahoo!知恵袋に投稿された質問文と回答文の計 60 個の文章に対して印象評価を行った．その結果，文章内容に関する因子が 9 個得られた．また，得られた因子の因子得点を適宜利用することで，通常の回答文に対し，「ベストアンサー」を特定できる可能性を示した[1][2].

しかし，ここで得られた因子得点は，評価実験を行った結果得られた質問文と回答文の文章 60 個に対するもののみであって，他の多数の質問文と回答文に対する因子得点は得られていない．そこで，どのような回答文に対しても「ベストアンサー」の推定を可能にすることを目的として，文章の特徴量から文章の因子得点を推定した結果について述べる．ここでは，重回帰分析を使用して，因子得点を推定する．文章の特徴量として，名詞や動詞などの品詞，ひらがな，カタカナ，英数字の出現回数あるいは比率など，形態素解析を使用して求められるものを採用している[3].

因子得点の推定実験を行い，2 因子はやや良好の推定精度が得られたが，残りの 7 因子は，十分な推定精度が得られないことが明らかとなった[3]．また，二次の項（説明変数同士の積）を考慮した重回帰分析を行い，いずれの因子も推定精度が向上すること，特に、推定精度が良好である因子が 3 因子，推定精度がやや良好である因子が 3 因子あることも明らかとした．しかしながら，残りの 3 因子に関しては，十分な推定精度が得られないことも明らかとなった[4].

これらの実験の基礎データとして使用した NTT データベースシリーズ[5][6]には，人が主観的に評定を行ったデータと，14 年間にわたる新聞記事に出現した単語や文字の出現回数を計測した客観的データも収録されている．これらのデータは，人間の言語処理過程に大きな影響を及ぼすものとして広く知られており，収録されている各特性値や特性値間の関係は，日本語自体の特性を示しているといえる[6]．これらのデータも文章の特徴量として有用であると考えられる．

そこで，本論文では，推定精度を更に向上させる方法として，単語心像性[5]を用いて，質問回答文の因子得点の推定精度の向上を試みる．分析の結果，9 因子中 2 因子には，分析結果に変動が見られた．

以降，2. では質問回答文に対する印象評価実験について述べ，3. ではこれまでの因子得点の推定について述べる．次に，4. で新たに特徴量として使用する単語心像性について述べ，5. で単語心像性を特徴量に加えて推定した結果について述べる．そして，6. では分析結果に対する考察を述べ，最後に 7. でまとめる．

2．質問回答文の印象評価実験

質問者と回答者の相性を判定することを目的として，質問者と回答者の文章の印象評価を行ってきた[1][2]．そこでは，50 個の印象語を使用し，Yahoo!知恵袋に投稿された質問文と回答文の計 60 個の文章に対して印象評価を行っている．

得られた評価値に対して因子分析を施した結果，的確性，不快性，独創性，容易性，執拗性，曖昧性，感動性，努力性，熱烈性という 9 個の因子が得られた．また，これらの因子の因子得点を利用することで，通常の回答文に対し，「ベストアンサー」を特定できる可能性を示してきた[1][2].

9 つの因子とそれぞれに対応する印象語をまとめて表 1 に示す.

† 京都工芸繊維大学 大学院工芸科学研究科, Graduate School of Information Science, Kyoto Institute of Technology
‡ 筑波大学大学院 図書館情報メディア研究科, Graduate School of Library, Information and Media Studies, University of Tsukuba

表 1.　9 つの因子と対応する印象語

因子	印象語		
第1因子（的確性）	説得力がある	流暢な	重要な
	素晴らしい	好ましい	巧みな
	真実味がある	清々しい	妥当な
	充実した	美しい	的確な
	丁寧な		
第2因子（不快性）	不快な	憤慨した	非常識な
	残念な	不当な	呆れる
	幻滅した	怖い	
第3因子（独創性）	独創的な	予想外な	特殊な
	斬新な	不思議な	
第4因子（容易性）	易しい	明瞭な	難しい
第5因子（執拗性）	細かい	しつこい	長い
第6因子（曖昧性）	曖昧な	不充分な	
第7因子（感動性）	心温まる	感動的な	
第8因子（努力性）	涙ぐましい		
第9因子（熱烈性）	熱い	力強い	

3. 因子得点の推定

3.1　文章の特徴量

文の長さや数，品詞の数を求めるために，Text Seer[7]を用いて形態素解析を行った．

文章において，複数回出現する単語が存在する可能性が高いことを考慮して，語彙数と語数とは独立した特徴量として抽出する．ここで，語彙数とは，文章中に同じ単語が複数回出現した場合でも 1 個と数えることを表し，語数とは，単純に単語の出現回数を表す．例えば，「私は私の道を切り開いていく」という文章を例にすると，「私」という単語が 2 回出現しているので，「私」という 1 つの語彙に対して，「私」の語数は 2 である．

また，ひらがな，漢字，カタカナ，記号，英数字に関しては，出現の割合によって印象が変わると考えられる．そこで，それぞれの出現回数だけではなく，それらを含む文章そのものの長さ及び文章内における含有率（全文字数に対する当該字種の文字数の比率）も考慮する必要があると考え，ひらがな，漢字，カタカナなどの文章における含有率も特徴量とした．例えば，表 2 の f_{16} の「ひらがな（語数）」は，文章内のひらがなの単語数を表し，f_{36} の「ひらがな（%）」は文章におけるひらがなの含有率を表している．

f_{26} の「未知語」とは，Text Seer をデフォルトの状態で使用し，「未知語」と判定された語数を表している．なお，未知語と判定された語は，名詞または記号として辞書に登録した上で，改めて形態素解析を行った．f_{40} の TTR（Type Token Ratio）は，語数に対する語彙数の比率を表している．

以上により，64 個の特徴量を使用することとした[3]．これを表 2 に示す．なお，$f_1, f_2, \cdots, f_{64}$ の表記は，後に重回帰式を表すために使用する．

表 2. 文章の特徴量（64 個）

f	特徴量	f	特徴量
f1	文字数	f33	感動詞（語数）
f2	名詞（語彙数）	f34	助動詞（語数）
f3	動詞（語彙数）	f35	助詞（語数）
f4	形容詞（語彙数）	f36	ひらがな（%）
f5	副詞（語彙数）	f37	漢字（%）
f6	連体詞（語彙数）	f38	カタカナ（%）
f7	接続詞（語彙数）	f39	記号（%）
f8	感動詞（語彙数）	f40	TTR
f9	助動詞（語彙数）	f41	全角記号(%)
f10	助詞（語彙数）	f42	英数字(%)
f11	接頭詞	f43	全角英数字(%)
f12	記号（語彙数）	f44	半角英数字(%)
f13	文数	f45	名詞(%)
f14	文の長さ平均（語数）	f46	動詞(%)
f15	文の長さ平均（字数）	f47	形容詞(%)
f16	ひらがな（語数）	f48	副詞(%)
f17	漢字（語数）	f49	連体詞(%)
f18	カタカナ（語数）	f50	接続詞(%)
f19	記号（語数）	f51	感動詞(%)
f20	全角記号（語数）	f52	助動詞(%)
f21	英数字（語数）	f53	助詞(%)
f22	全角英数字（語数）	f54	「!」の数
f23	半角英数字（語数）	f55	「?」の数
f24	語数	f56	句点の数
f25	語彙数	f57	読点の数
f26	未知語	f58	中点の数
f27	名詞（語数）	f59	3点リーダの数
f28	動詞（語数）	f60	鍵括弧の数
f29	形容詞（語数）	f61	鍵括弧閉の数
f30	副詞（語数）	f62	括弧の数
f31	連体詞（語数）	f63	括弧閉の数
f32	接続詞（語数）	f64	「/」の数

3.2　多重共線性の考慮

重回帰分析を実施する際は，複数の説明変数同士は無相関であるという前提が必要となり，説明変数は以下の条件を考慮して選択しなければならない．

a) 目的変数との相関係数が高い説明変数の選択

b) 高い相関を示す説明変数の組のうち，一方を説明変数から除外

ここで，b)の事項に反すると，偏回帰係数が正しく求まらないことがあり，このような状態を多重共線性という．多重共線性を確認するには，「目的変数との相関係数」と「回帰係数」との符号が逆転している説明変数を調べる方法がある[8]．符号が一致しない原因は，説明変数の組の中に高い相関のある説明変数が含まれているからである．

多重共線性を回避するために，表 2 に示す説明変数に関して，説明変数同士の相関係数の値を調べ，0.7 以上である組に関しては，一方を説明変数から除外した．その結果，説明変数は 39 個となった．これを表 3 に示す．

表 3. 多重共線性を考慮した説明変数（39 個）

f	変数	f	変数
f1	文字数	f41	全角記号(%)
f9	助動詞（語彙数）	f42	英数字(%)
f11	接頭詞	f43	全角英数字(%)
f12	記号（語彙数）	f45	名詞(%)
f13	文数	f46	動詞(%)
f15	文の長さ平均（字数）	f47	形容詞(%)
f18	カタカナ（語数）	f48	副詞(%)
f19	記号（語数）	f49	連体詞(%)
f20	全角記号（語数）	f50	接続詞(%)
f22	全角英数字（語数）	f51	感動詞(%)
f29	形容詞（語数）	f54	「!」の数
f30	副詞（語数）	f55	「?」の数
f31	連体詞（語数）	f56	句点の数
f32	接続詞（語数）	f57	読点の数
f33	感動詞（語数）	f58	中点の数
f36	ひらがな（%）	f59	3点リーダの数
f37	漢字（%）	f60	鍵括弧の数
f38	カタカナ（%）	f62	括弧の数
f39	記号（%）	f64	「/」の数
f40	TTR		

3.3 単項のみを考慮した推定結果

2. で述べた 9 つの因子の因子得点を，それぞれ $y_1, y_2, \cdots, y_9$ と定める．ここでは，2. の印象評価実験で使用した 60 個の質問回答文に対して，文章の特徴量を説明変数，因子得点を目的変数として，ステップワイズ選択法[10]による重回帰分析を行った．

この結果，重回帰式(1)が得られた．但し，第 9 因子に関しては，重回帰式が得られなかった．

$$
\begin{cases}
y_1 = 0.00579 f_1 - 0.131 f_{22} + 0.0851 f_{29} + 0.484 f_{62} \\
\quad + 0.0526 f_{43} - 0.0147 f_{38} + 0.121 f_{60} + 0.0101 f_{45} \\
\quad - 0.0740 f_{57} - 0.00228 f_{15} - 0.582 \\
y_2 = -0.0938 f_9 + 0.369 \\
y_3 = -0.0845 f_9 + 0.0444 f_{22} + 0.245 \\
y_4 = -0.00348 f_1 + 0.0588 f_{22} - 0.140 f_{55} + 0.0673 f_9 \\
\quad + 0.181 \\
y_5 = 0.0229 f_1 + 0.108 f_9 + 0.0978 f_{55} + 0.0169 f_{18} \\
\quad + 0.0142 f_{37} + 0.464 f_{59} - 0.0232 f_{19} - 0.0958 f_{47} \\
\quad + 0.161 f_{49} + 0.192 f_{11} - 1.02 \\
y_6 = -0.00340 f_1 - 0.00828 f_{36} + 0.689 \\
y_7 = 0.456 f_{33} + 0.0836 f_{30} + 0.102 f_{47} - 0.0330 f_{56} \\
\quad - 0.193 \\
y_8 = 0.108 f_9 - 0.00628 f_{42} - 0.00120 f_1 + 0.0826 f_{47} \\
\quad - 0.305
\end{cases}
\tag{1}
$$

重相関係数と，選ばれた説明変数を，それぞれ表 4，表 5 に示す．重相関係数は，その値が 0.9 以上ならば，分析精度が非常に良好であるとされ，0.7 以上ならば，分析精度がやや良好であるとされ，0.7 未満ならば，分析精度が不良であるとされている[9]．

表4. 単項のみを考慮した場合の重相関係数

因子	重相関係数
第1因子（的確性）	0.879
第2因子（不快性）	0.350
第3因子（独創性）	0.475
第4因子（容易性）	0.643
第5因子（執拗性）	0.905
第6因子（曖昧性）	0.677
第7因子（感動性）	0.562
第8因子（努力性）	0.587

表5. 単項のみを考慮した場合に選択された各因子の説明変数

因子	説明変数
第1因子（的確性）	文字数 全角英数字(語数) 形容詞（語数） 括弧の数 全角英数字(%) カタカナ（%） 鍵括弧の数 名詞（%） 読点の数 文の長さ平均（字数）
第2因子（不快性）	助動詞（語彙数）
第3因子（独創性）	助動詞（語彙数） 全角英数字（語数）
第4因子（容易性）	文字数 全角英数字（語数） 「?」の数 助動詞（語彙数）
第5因子（執拗性）	文字数 助動詞（語彙数） 「?」の数 カタカナ（語数） 漢字（%） 3点リーダの数 記号（語数） 形容詞(%) 連体詞(%) 接頭詞
第6因子（曖昧性）	文字数 ひらがな（%）
第7因子（感動性）	感動詞（語彙数） 副詞（語彙数） 形容詞(%) 句点の数
第8因子（努力性）	助動詞（語彙数） 英数字(%) 文字数 形容詞(%)

表 4 の結果から，第 5 因子（執拗性）は，0.9 以上の値であるので，分析精度が非常に良好であるといえる．また，第 1 因子（的確性）は，0.7 以上の値であるから，分析精度はやや良好であるといえる．一方，その他の 7 因子は 0.7 未満の値であり，分析精度は良好とは言えない．また，第 9 因子は，該当する説明変数が得られなかった．

$$
\begin{cases}
y_1 = 0.00121 f_1 f_9 - 0.00814 f_{15} f_{31} + 0.0756 f_{30} f_{47} + 0.00307851 f_{38} f_{39} + 0.153 f_{29} f_{60} - 0.0196 f_{19} f_{30} - 0.00885 f_{38} f_{48} + 0.772 f_{40} \\
\quad - 1.64 f_{11} f_{33} - 0.00512 f_{12} f_{37} + 0.299 f_{56} f_{59} - 0.110 f_{47} f_{55} + 0.205 f_{31} f_{47} - 0.0338 f_{15} f_{51} + 0.00310 f_1 f_{51} + 0.00620 f_{22} f_{45} \\
\quad + 0.0262 f_{41} f_{49} - 0.0347 f_{31} f_{42} + 0.0345 f_{13} f_{62} - 0.00828 f_{38} + 0.00846 f_{30} f_{39} - 0.0143 f_{43} f_{56} + 0.0229 f_{31} f_{38} + 0.0172 f_{18} f_{50} \\
\quad + 0.0712 f_{31} f_{55} + 0.0738 f_{54} f_{56} - 0.0462 f_{32} f_{47} + 0.0245 f_{39} f_{40} + 0.000212 f_{15} f_{38} - 0.00125 f_{15} f_{55} - 0.961 \\
y_2 = -0.00193 f_9 f_{36} - 0.0725 f_{39} f_{40} + 1.13 f_{33} f_{54} + 0.698 \\
y_3 = 1.33 f_{40} f_{40} + 0.000302 f_1 f_{22} - 0.0143 f_{38} f_{58} - 0.00351 f_{29} f_{37} + 0.0126 f_{31} f_{46} - 0.636 \\
y_4 = -0.000836 f_1 f_9 - 2.19 f_{49} f_{51} + 0.00390 f_9 f_{19} - 0.0220 f_{37} f_{40} + 0.00248 f_9 f_{37} + 0.00109 f_{38} f_{42} + 0.00487 f_{29} f_{38} + 0.382 \\
y_5 = 0.000544 f_1 f_9 + 1.07 f_{49} f_{59} + 0.00257 f_{18} f_{46} - 0.0226 f_{18} f_{50} + 0.159 f_{49} + 0.344 f_{12} f_{51} + 0.00273 f_9 f_{18} - 0.0370 f_{43} \\
\quad + 0.0896 f_{47} f_{58} - 0.0317 f_{31} f_{31} - 0.113 f_{45} f_{51} - 0.115 f_{50} - 0.352 f_{33} f_{40} + 0.00675 f_{45} f_{55} + 0.00448 f_{42} - 0.00526 f_{45} f_{60} \\
\quad + 0.312 f_{31} f_{33} - 0.000530 f_{19} f_{19} - 0.00568 f_{39} f_{47} - 0.439 \\
y_6 = 0.888 f_{40} f_{40} + 0.000318 f_{36} f_{37} - 0.00351 f_{37} f_{57} - 0.00334 f_{37} f_{43} - 0.0923 f_9 f_{40} + 0.0837 f_{54} f_{56} + 0.345 f_{11} f_{49} - 0.314 \\
y_7 = 0.469 f_{33} + 0.147 f_{30} - 0.0184 f_{55} f_{56} + 0.231 f_{31} f_{54} - 0.0150 f_{57} f_{58} - 0.0208 f_{48} f_{57} - 0.141 \\
y_8 = 0.162 f_9 f_{40} + 0.0211 f_{31} f_{39} - 0.00493 f_{42} - 0.749 f_{33} f_{50} + 0.237 f_{32} f_{59} - 0.0355 f_{12} f_{47} + 0.0123 f_{41} f_{48} - 0.0886 f_9 f_{51} - 0.372
\end{cases} \quad (2)
$$

3.4 二次の項を考慮した場合の推定結果

ここでは，二次の項（説明変数同士の積）を考慮した重回帰分析を行う．2.の印象評価実験で使用した 60 個の質問回答文に対して，表 3 に示す文章の特徴量を説明変数，因子得点を目的変数として，重回帰分析を行った．

この結果，重回帰式(2)が得られた．ここでも，第 9 因子に関しては，重回帰式が得られなかった．

また，この時の目的変数と説明変数の重相関係数を表 6 に示す．

表 6. 二次の項を考慮した場合の重相関係数

因子	重相関係数
第1因子（的確性）	0.997
第2因子（不快性）	0.618
第3因子（独創性）	0.716
第4因子（容易性）	0.844
第5因子（執拗性）	0.984
第6因子（曖昧性）	0.860
第7因子（感動性）	0.711
第8因子（努力性）	0.772

表 4 と比較すると，表 6 の方がどの因子も重相関係数の値が向上していることがわかる．従って，単項のみを考慮した重回帰分析よりも，二次の項も考慮した重回帰分析の方が，分析精度が向上していることがわかる．

各因子の重相関係数に関して，第 1 因子（的確性），第 5 因子（執拗性）の 2 因子は，0.9 以上の値をとっているため，分析精度が非常に良好であるといえる．また，第 3 因子（独創性），第 4 因子（容易性），第 6 因子（曖昧性），第 7 因子（感動性），第 8 因子（努力性）の 5 因子に関しては，0.7 以上の値をとっているので，分析精度がやや良好であるといえる．

一方，第 2 因子（不快性）は 0.7 未満の値であり，分析精度は不良であるといえる．第 9 因子（熱烈性）は，該当する説明変数がここでも得られなかった．

4. 単語心像性

NTT データベースシリーズ[5]には，人が主観的に評定を行ったデータと，14 年間にわたる新聞に単語や文字が出現した回数を数えた客観的データが収録されている．これらのデータは，人間の言語処理過程に大きな影響を及ぼすものとして広く知られており，収録されている各特性値や特性値間の関係は，日本語自体の特性を示しているといえる[6]．これらのデータも文章の特徴量として有用であると考えられる．

これらのデータの中でも，単語心像性を文章の特徴量に追加する．単語心像性とは，単語から喚起される様々なイメージが，どの程度思い浮かべやすいかを示す主観的特性である．例えば，「りんご」という言葉を聞くと，赤・黄・緑の丸い形の果物，甘くみずみずしい味・匂い，サクッとした音や歯ざわり，持った時の感触を思い浮かべることができる．一方，「世界」「経済」は，「りんご」に比べると具体的なイメージを思い浮かべにくいと思われる．

ここでは，単語心像性の特性値は，「単語の非言語的感覚イメージの喚起力」に関して，「1：イメージを非常に思い浮かべにくい（または思い浮かばない）～7：イメージを非常に思い浮かべやすい」の 7 段階尺度で評定させた値である．新聞記事を対象としたデータ[5]と，質問回答文の文章を形態素解析したデータとを比較して，収録データに合致する単語が形態素解析したデータに存在するならば，その単語の単語心像性の値を特性値として使用する．なお，形態素解析したデータに収録データと単語が合致しない場合は，その単語の単語心像性の値は考慮しないものとして処理する．

また，単語の同じ表記でも，意味または読みが異なる場合がある．例えば，意味が異なる例としては，「アース」という単語は，「電気を逃がすために接地すること」，「地球」，「殺虫剤（メーカー）」の意味がある．読みが異なる例としては，「間」という言葉は，「あいだ」，「ま」の読みがある．このような単語が形態素解析したデータに存在する場合は，文脈から判断しながら手動で意味または読みを決定する．

このようにして，単語心像性の特徴量を抽出した．これを表 7 に示す．特徴量としては，単語心像性に該当した単語の数や該当した単語の割合や，単語心像性の値が 1 点台，2 点台……のように，1 点間隔で特徴量をとったものや，1.0 以上 1.5 未満，1.5 以上 2.0 未満，……のように，0.5 点間隔で特徴量をとったもの，を採用した．

表 7.　単語心像性の特徴量

該当単語（語彙数）	1点台（語数）
該当単語（語数）	1.0～1.5未満（語数）
該当単語率（語数）	1.5～2.0未満（語数）
1点台（語彙数）	2点台（語数）
1.0～1.5未満（語彙数）	2.0～2.5未満（語数）
1.5～2.0未満（語彙数）	2.5～3.0未満（語数）
2点台（語彙数）	3点台（語数）
2.0～2.5未満（語彙数）	3.0～3.5未満（語数）
2.5～3.0未満（語彙数）	3.5～4.0未満（語数）
3点台（語彙数）	4点台（語数）
3.0～3.5未満（語彙数）	4.0～4.5未満（語数）
3.5～4.0未満（語彙数）	4.5～5.0未満（語数）
4点台（語彙数）	5点台（語数）
4.0～4.5未満（語彙数）	5.0～5.5未満（語数）
4.5～5.0未満（語彙数）	5.5～6.0未満（語数）
5点台（語彙数）	6点台（語数）
5.0～5.5未満（語彙数）	6.0～6.5未満（語数）
5.5～6.0未満（語彙数）	6.5～7.0未満（語数）
6点台（語彙数）	
6.0～6.5未満（語彙数）	
6.5～7.0未満（語彙数）	

3.で列挙した文章の特徴量と同様に，多重共線性を回避するために，表 7 に示す特徴量間同士に関して，それぞれの相関係数を調べた．その結果，特徴量間同士の相関係数のほとんどが 0.7 以上となり，各特徴量の間に強い相関が見られることがわかった．このうち，相関係数が 0.7 未満の組である「4 点台（語数）」，「6.5～7.0 未満（語数）」を特徴量として使用することにする．ここでは，それぞれを f_{65}, f_{66} とする．

5. 単語心像性を加えた推定

5.1　単項のみを考慮した場合

2.で使用した 60 個の質問回答文に対して，表 3 に示す 39 個の特徴量に，4.で使用した 2 個の特徴量を追加し，計 41 個の特徴量を説明変数，因子得点を目的変数として，ステップワイズ選択法による重回帰分析を行った．

$$
\begin{cases}
y_1 = 0.0287 f_{65} + 0.195 f_{32} + 0.0541 f_{43} + 0.0118 f_{20} \\
\quad - 0.0127 f_{38} - 0.155 f_{55} + 0.0794 f_{29} - 0.426 \\
y_2 = -0.0938 f_9 + 0.369 \\
y_3 = -0.0845 f_9 + 0.0444 f_{22} + 0.245 \\
y_4 = -0.00503 f_{65} + 0.0828 f_{29} - 0.161 f_{55} + 0.0583 f_9 \\
\quad + 0.214 f_{54} + 0.103 \\
y_5 = 0.00229 f_1 + 0.108 f_9 + 0.0978 f_{55} + 0.0169 f_{18} \\
\quad + 0.0142 f_{37} + 0.464 f_{59} - 0.0232 f_{19} - 0.0958 f_{47} \\
\quad + 0.161 f_{49} + 0.192 f_{11} - 1.02 \\
y_6 = -0.00340 f_1 - 0.00828 f_{36} + 0.926 \\
y_7 = 0.456 f_{33} + 0.0836 f_{30} + 0.102 f_{47} - 0.0330 f_{56} \\
\quad - 0.193 \\
y_8 = 0.108 f_9 - 0.00628 f_{42} - 0.00120 f_1 + 0.0826 f_{47} \\
\quad - 0.305
\end{cases}
\tag{3}
$$

この結果，重回帰式(3)が得られた．ここでも，第 9 因子に関しては，重回帰式が得られなかった．

また，重相関係数と，選ばれた説明変数を，それぞれ表 8，表 9 に示す．重相関係数に関して，表 4 と表 8 とを比較すると，第 1 因子（的確性）の分析精度が低下し，第 4 因子（容易性）の分析精度が向上している．しかし，他の因子に関しては，分析精度に全く変化が無かった．また，表 5 と表 9 とを比較すると，選ばれた説明変数も，重相関係数と同様に，第 1 因子と第 4 因子を除いては，全く説明変数に変化が無かった．

表8. 単語心像性の特徴量を追加し単項のみを考慮した場合の重相関係数

因子	重相関係数
第1因子（的確性）	0.809
第2因子（不快性）	0.350
第3因子（独創性）	0.475
第4因子（容易性）	0.737
第5因子（執拗性）	0.905
第6因子（曖昧性）	0.677
第7因子（感動性）	0.562
第8因子（努力性）	0.587

表 9. 単語心像性の特徴量を追加し単項のみを考慮した場合に選択された説明変数

因子	説明変数
第1因子（的確性）	単語心像性4点台
	接続詞（語数）
	全角英数字(%)
	全角記号（語数）
	カタカナ（%）
	「?」の数
	形容詞（語数）
第2因子（不快性）	助動詞（語彙数）
第3因子（独創性）	助動詞（語彙数）
	全角英数字（語数）
第4因子（容易性）	単語心像性4点台
	形容詞（語数）
	「?」の数
	助動詞（語彙数）
	「!」の数
第5因子（執拗性）	文字数
	助動詞（語彙数）
	「?」の数
	カタカナ（語数）
	漢字（%）
	3点リーダの数
	記号（語数）
	形容詞(%)
	連体詞(%)
	接頭詞
第6因子（曖昧性）	文字数
	ひらがな（%）
第7因子（感動性）	感動詞（語彙数）
	副詞（語彙数）
	形容詞(%)
	句点の数
第8因子（努力性）	助動詞（語彙数）
	英数字(%)
	文字数
	形容詞(%)

$$
\begin{cases}
y_1 = 0.103 f_{40} f_{65} + 0.01350 f_{12} f_{20} - 0.0120 f_{18} f_{30} - 0.299 f_{11} f_{22} + 0.00598 f_{30} f_{45} + 0.0310 f_{11} f_{42} - 0.0695 f_{58} f_{60} - 0.00478 f_9 f_{20} \\
\quad + 0.0381 f_{43} + 0.0104 f_{45} - 0.00891 f_{38} - 0.0446 f_{30} - 0.00714 f_{15} f_{51} - 0.694 \\
y_2 = -0.00193 f_9 f_{36} - 0.0725 f_{39} f_{40} + 1.13 f_{33} f_{54} + 0.698 \\
y_3 = 1.33 f_{40} f_{40} + 0.000302 f_1 f_{22} - 0.0143 f_{38} f_{58} - 0.00351 f_{29} f_{37} + 0.0126 f_{31} f_{46} - 0.636 \\
y_4 = -0.000209 f_1 f_{65} - 2.46 f_{49} f_{51} - 0.0133 f_{37} f_{40} + 0.00111 f_{19} f_{65} + 0.0136 f_{38} f_{40} + 0.000265 f_1 f_{29} + 0.284 \\
y_5 = 0.000544 f_1 f_9 + 1.07 f_{49} f_{59} + 0.00257 f_{18} f_{46} - 0.0226 f_{18} f_{50} + 0.159 f_{49} + 0.344 f_{12} f_{51} + 0.00273 f_9 f_{18} - 0.0370 f_{43} \\
\quad + 0.0896 f_{47} f_{58} - 0.0317 f_{31} f_{31} - 0.113 f_{45} f_{51} - 0.115 f_{50} - 0.352 f_{33} f_{40} + 0.00675 f_{45} f_{55} + 0.00448 f_{42} - 0.00526 f_{45} f_{60} \\
\quad + 0.312 f_{31} f_{33} - 0.000530 f_{19} f_{19} - 0.00568 f_{39} f_{47} - 0.439 \\
y_6 = 0.888 f_{40} f_{40} + 0.000318 f_{36} f_{37} - 0.00351 f_{37} f_{57} - 0.00334 f_{37} f_{43} - 0.0923 f_9 f_{40} + 0.0837 f_{54} f_{56} + 0.345 f_{11} f_{49} - 0.314 \\
y_7 = 0.469 f_{33} + 0.147 f_{30} - 0.0184 f_{55} f_{56} + 0.231 f_{31} f_{54} - 0.0150 f_{57} f_{58} - 0.0208 f_{48} f_{57} - 0.141 \\
y_8 = 0.162 f_9 f_{40} + 0.0211 f_{31} f_{39} - 0.00493 f_{42} - 0.749 f_{33} f_{50} + 0.237 f_{32} f_{59} - 0.0355 f_{12} f_{47} + 0.0123 f_{41} f_{48} - 0.0886 f_9 f_{51} - 0.372
\end{cases} \quad (4)
$$

5.2 二次の項を考慮した場合

単項の場合と同様に，2.の印象評価実験で使用した 60 個の質問回答文に対して， 41 個の文章の特徴量を説明変数，因子得点を目的変数として，ステップワイズ選択法による重回帰分析を行った．

この結果，重回帰式(4)が得られた．ここでも，第 9 因子に関しては，重回帰式が得られなかった．

また，この時の重相関係数を表 10 に示す．

表10. 単語心像性の特徴量を追加し二次の項を考慮した場合の重相関係数

因子	重相関係数
第1因子（的確性）	0.943
第2因子（不快性）	0.618
第3因子（独創性）	0.716
第4因子（容易性）	0.854
第5因子（執拗性）	0.984
第6因子（曖昧性）	0.860
第7因子（感動性）	0.711
第8因子（努力性）	0.772

重相関係数に関して，表 6 と表 10 を比較すると，第 1 因子（的確性）の分析精度が低下し，第 4 因子（容易性）の分析精度がわずかに向上している．しかし，他の因子に関しては，分析精度に全く変化が無かった．また，選ばれた説明変数に関しても，重回帰式(2)と重回帰式(4)を比較すると，重相関係数と同様に，第 1 因子と第 4 因子を除いては，全く説明変数に変化が無かった．

6. 考察

単項のみを考慮した場合，各因子の重相関係数に関して，第 5 因子（執拗性）の因子のみは，分析精度が良好であり，的確性，容易性の 2 因子は分析精度がやや良好であるという結果が得られた．一方，残りの 6 因子は分析精度が不良であるという結果が得られた．単語心像性の特徴量を追加したことにより，わずかながら分析精度に変動が見られる．

また，二次の項を考慮した場合，各因子の重相関係数に関して，第 1 因子（的確性），第 5 因子（執拗性）の 2 因子は，0.9 以上の値をとっているため，分析精度が非常に良好であるといえる．また，第 3 因子（独創性），第 4 因子（容易性），第 6 因子（曖昧性），第 7 因子（感動性），第 8 因子（努力性）の 5 因子は，0.7 以上の値をとっているので，分析精度がやや良好であるといえる．一方，第 2 因子（不快性）は 0.7 未満の値であり，第 9 因子（熱烈性）は該当する説明変数が得られなかったため，分析精度は不良であるといえる．

回帰式についての考察を行うために，各因子の標準偏回帰係数のうち，絶対値が大きいものを表 11 に示す．

表 11. 標準偏回帰係数の絶対値の大きいもの

因子	変数	係数
第1因子（的確性）	f12*f20	1.05
	f40*f65	0.602
第2因子（不快性）	f9*f36	0.618
第3因子（独創性）	f1*f22	0.605
	f40*f40	0.562
第4因子（容易性）	f1*f65	-1.09
第5因子（執拗性）	f1*f9	0.984
	f12*f51	0.655
第6因子（曖昧性）	f37*f57	-0.511
第7因子（感動性）	f30	0.691
第8因子（努力性）	f9*f40	0.506

第 1 因子（的確性）では， $f_{12} * f_{20}$ の係数が正で大きい．従って，記号の語彙数（すなわち，記号の種類）が多く，かつ，全角記号の語数が多いほど，的確性の因子得点が高くなると考えられる．また， $f_{40} * f_{65}$ の正の係数も大きい．TTR の値が大きく（すなわち，同じ語彙が繰り返し使用されにくい），かつ，単語心像性 4 点台（すなわち，単語からのイメージがある程度思い浮かべやすいもの）の単語が多いほど，的確性の因子得点が高くなると考えられる．

第 2 因子（不快性）では， $f_9 * f_{36}$ の係数が正である．従って，助動詞の語彙数が多く，ひらがなの割合が少ない場合は，不快性の因子得点が高くなると考えられる．

第 3 因子（独創性）では， $f_1 * f_{22}$ の係数が正である．文字数が多く，かつ，全角英数字の語数が多いほど，独創性の因子得点が高くなると考えられる．また， $f_{40} * f_{40}$ の係数も大きい．TTR の値が大きいと，独創性の因子得点が高くなると考えられる．

第 4 因子（容易性）では，$f_1 * f_{65}$ の係数は負である．この場合，文字数と単語心像性 4 点台の単語の，一方が多く，他方が少ないほど，容易性の因子得点が高くなると考えられる．

第 5 因子（執拗性）では，$f_1 * f_9$ の係数が大きい．文字数が多く，かつ，助動詞の語彙数が多いほど，執拗性の因子得点が高くなると考えられる．また，$f_{12} * f_{51}$ の係数も大きい．記号の語彙数が多く，かつ，感動詞の割合が高いほど，執拗性の因子得点が高くなると考えられる．

第 6 因子（曖昧性）では，$f_{37} * f_{57}$ の係数は負である．漢字の割合が少なく，かつ，読点の数が少ないほど，曖昧性の因子得点が高くなると考えられる．

第 7 因子（感動性）では，f_{30} の正の係数が大きい．副詞の語数が多いほど，感動性の因子得点が高くなると考えられる．

第 8 因子（努力性）では，$f_9 * f_{40}$ の係数は正である．助動詞の語彙数が多く，かつ，TTR の値が大きいほど，努力性の因子得点が高くなると考えられる．

7. まとめ

本論文では，質問者と回答者の相性を判定することを目的として，文章の因子得点の推定精度の向上をめざして検討を行った．ここでは，単語心像性[5]を特徴量に追加して，質問回答文の因子得点の推定精度の向上を試みた．その結果，9 因子中 2 因子のみではあるが，分析結果に変動が見られた．

依然として，第 9 因子については重回帰式が得られていない．また，単語心像性を文章の特徴量に加えたが，第 1 因子，第 4 因子，第 9 因子を除く 6 因子に関しては，分析結果に全く影響を及ぼしていない．今回評価に追加した単語心像性以外にも，単語親密度，表記妥当性など[5]，文章の印象に影響すると考えられる特徴量が知られていることから，今後，これらを特徴量に加えた上で，再度重回帰分析を行い，第 9 因子の重回帰式を求めてゆく予定である．また，さらなる因子得点の推定精度の向上を図ることが課題である．

謝辞

本研究は一部，科研費（21500091）の助成を受けて行われたものである．また，実装・評価に際し，大学共同利用機関法人国立情報学研究所から提供を受けた，Yahoo!知恵袋のデータを利用している．ここに記して謝意を示す．

参考文献

[1] 横山友也，宝珍輝尚，野宮浩揮，佐藤哲司：質問回答サイトの質問文と回答文の印象評価，第 2 回データ工学と情報マネジメントに関するフォーラム（DEIM2010），C4-2，2010.

[2] 横山友也，宝珍輝尚，野宮浩揮，佐藤哲司：質問回答サイトの質問文と回答文の印象評価とベストアンサーの推定，日本感性工学会論文誌，Vol.10，No.2，pp.221-230，2011.

[3] 横山友也，宝珍輝尚，野宮浩揮，佐藤哲司：文章の特徴量を用いた質問回答文の因子得点の推定，第 6 回日本感性工学春季大会，22D-2，2011.

[4] 横山友也，宝珍輝尚，野宮浩揮，佐藤哲司：文章の特徴量を用いた質問回答文の因子特典の推定精度の向上，日本感性工学会関西支部大会 2011

[5] 佐久間尚子，伊集院睦雄，伏見貴夫，辰巳格，田中正之，天野成昭，近藤公久：単語心像性①，NTT データベースシリーズ日本語の語彙特性 第 3 期（第 8 巻），（社）三省堂，2005.

[6] NTT データベースシリーズ，
http://www.kecl.ntt.co.jp/mtg/goitokusei/

[7] Text Seer マニュアル
http://www.valdes.titech.ac.jp/~t_kawa/ts/manual.htm

[8] 菅民郎，初心者がらくらく読める多変量解析の実践上，pp.37-41，（社）現代数学社，1993.

[9] 菅民郎，初心者がらくらく読める多変量解析の実践上，pp.42-45，（社）現代数学社，1993.

[10] 重回帰分析（ステップワイズ変数選択），
http://aoki2.si.gunma-u.ac.jp/R/sreg.html