文章の特徴量を用いた質問回答文の印象の因子得点の推定精度の向上

Improving Estimation Accuracy of Factor Scores of Impression of Question and Answer Statements by Using Feature Values of Statements

横山 友也*, 宝珍 輝尚*, 野宮 浩揮*, 佐藤 哲司**

Yuya Yokoyama, Teruhisa Hochin, Hiroki Nomiya, Tetsuji Satoh

*京都工芸繊維大学 情報工学専攻, 京都市左京区松ヶ崎御所海道町

Kyoto Institute of Technology, Graduate School of Information Science

Goshokaidocho, Matsugasaki, Sakyo-ku, Kyoto

**筑波大学大学院 図書館情報メディア研究科, 茨城県つくば市春日 1-2

Graduate School of Library, Information and Media Studies, University of Tsukuba

1-2 Kasuga, Tsukuba-shi, Ibaraki

あらまし:質問回答サイトにおける質問者と回答者のミスマッチングの問題を解消するために, Yahoo!知恵袋に投稿された 60 個の質問回答文に対し, 印象評価実験を行ってきた. 実験結果に対して因子分析を施したところ, 文章に関する 9 個の因子が得られた. 構文情報, 単語心像性, 文末表現といった特徴量を採用することで, 全ての質問回答文の因子得点の推定を試みてきている. 本論文では, 単語親密度や表記妥当性を特徴量として採用することにより, 推定精度の向上を図る. これらの特徴量を追加することで, 推定精度が向上できることを示す.

Summary: In order to avoid the problem of mismatch between the questioner and the respondent at Q & A sites, we have experimentally evaluated the impression of 60 question and answer statements posted at Yahoo! Chiebukuro. Nine factors as to statements have been obtained by applying the factor analysis to the scores obtained through the experiment. Factor scores of any other statements have been tried to be estimated by adopting such feature values as syntactic information, word imageability and closing sentence expressions. This paper tries to improve estimation accuracy by adopting word familiarity and notation validity as feature values of statements. It is shown that estimation accuracy can be improved by adding these feature values to the conventional ones.

キーワード:Q&A サイト, 重回帰分析, 因子得点, 単語親密度, 表記妥当性

Keywords: Question & Answer site, multiple regression analysis, factor score, word familiarity, notation validity

1. はじめに

インターネット上における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形態であり, 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なってきている[1]. あるユーザが質問を投稿すると, 他のユーザ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．

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[2]．Yahoo!知恵袋[1]に投稿された質問文と回答文の計60個の文章に対して，50個の印象語を使用して，印象評価を行った．得られた評価値に対して因子分析を行った結果，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，「ベストアンサー」を特定できる可能性を示した[2]．しかし，ここで得られた因子得点は，評価実験を行っ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，どのような質問文と回答文に対しても因子得点の推定を可能にすることを目的として，文章の特徴量からの文章の因子得点の推定を重回帰分析により試みてきた[3]．ここで，(ⅰ)形態素解析を通して得られた「構文情報」，(ⅱ)「単語心像性」，(ⅲ)文末表現という3種の特徴量を用いて因子得点を推定すると，7因子は良好に，2因子はやや良好に因子得点が推定できることを示してきた[3]．しかし，やや良好な推定精度が得られた2因子に関しては，更に推定精度を良くする余地があり，文章の特徴量を更に追加することで，因子得点の推定精度をより向上させる可能性がある．

そこで，本論文では，文章の特徴量として，単語親密度と表記妥当性を追加することにより因子得点の推定精度の向上を試みる．これらの特徴量を用いて推定を行い，推定精度が向上することを示す．

以降，2.では関連研究について述べ，3.ではこれまでの研究について述べる．4.では，新たに追加した特徴量について述べ，5.では因子得点の推定結果について述べる．そして，6.で推定結果に対する考察を示す．最後に7.でまとめる．

2. 関連研究

これまで，「ベストアンサー」を推定する研究が行われてきている[4–11]．Bloomaらは，非テキスト特徴量とテキスト特徴量を用いて，「ベストアンサー」の推定を試みている[4]．Agichteinらは，内容や語法の特徴量を使用することによって，質問文と回答文の質の評価を試みている[5]．また，類推による手法も提案されている[6]．この手法では，過去の知見における質問文と回答文のリンクを使用することによって，「ベストアンサー」を探索する．Kimらは，「ベストアンサー」の選択基準を提案している[7]．情報型の質問は，文章内容の特徴量が重要であり，提案型の質問には有用性が重要であり，選択型の質問には社会的な感情が重要であるとしている．

西原らは，ある1つの質問文に対する回答文群より，「ベストアンサー」になりやすいものを検出する手法を提案している[11]．質問者と回答者の文末表現の相性に着目し，質問文と「ベストアンサー」の組み合わせをクラスタリングすることで，一定の成果をあげている．しかし，この研究では，質問文ならびに回答文の文末表現に着目した手法をとっており，文章内容に着目した手法をとっていない．そこで，本研究では，文末表現も，特徴量として考慮した上で，文章全体の文体や内容から受ける印象評価に着目して検討を行っている．

また，これらの研究は「ベストアンサー」を推定することに重点が置かれた研究である．この点に対して，本研究では，質問文に適切な回答を施すことができると考えられる回答者の選出を図っている点が，大きな特色となっている．更に，既存の研究では，質問回答文の構文情報を直接扱うことにより，「ベストアンサー」の推定を図っている．この点に対して，本研究では，同じ内容の文章でも，表現によって受ける印象が大きく異なることを考慮しているため，質問回答文の印象の度合いにも着目している点にも，大きな特色があるといえる．

一方，熊本は新聞記事を対象として印象評価を行っている[17]．被験者100人が新聞記事10記事を読んで，印象語42語のそれぞれについて5段階(強い—わりと強い—わりと弱い—弱い—なし)で評価するという印象評価実験(アンケート調査)を9度実施して，印象評価データ(印象語42語×9000件)を収集・分析することによって，新聞記事の印象を表現するのに適した印象軸を提案している[17]．本研究では，質問回答文を対象として，印象語を用いた印象評価実験を行う．

3. これまでの研究

3.1. 印象語

Yahoo!知恵袋[1]に実際に投稿された12組60個の4大ジャンル(Yahoo!オークション，パソコン・周辺機器，恋愛相談・人間関係，政治・社会問題)の質問回答文に対して，印象評価実験を行い，実験結果に対して因子分析を施したところ，文章に関する因子が9個得られた[2]．因子，ならびに，因子に対応する印象語を表1に示す．なお，これらの因子の因子得点は，実験で使用した60個の質問回答文から得ており，実験で使用していない質問回答文の因子得点はまだ求まっていない．そこで，文章の特徴量から因子得点を重回帰分析による推定を試みた．

表1　9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	

3.2. 文章の特徴量

文章の特徴量として，形態素解析を通して得られた構文情報，単語心像性，文末表現の3組を採用してきた[3].構文情報とは，文の数や文の長さ，名詞や動詞といった品詞の数や割合を，形態素解析により抽出した．また，感嘆符や疑問符の数といった具体的な記号も，文章の特徴量に採用している．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を意味している.また，文末表現とは，「全ての質問文と回答文に含まれる文末表現」を表している.ここでは，「ぞ」「だ」「よ」「ね」等の助詞の語数・割合や，文末にくる語数・割合を採用している．

ところで，重回帰分析を実施する際，複数の説明変数同士は無相関であるという前提が必要であり，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[12].

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一方を説明変数から除外

ここで，b)の事項に反した場合，偏回帰係数が正しく求まらないことがあり，この状態を多重共線性という[12]．多重共線性を回避するために，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，一方のみを説明変数として使用し，他方を除外する必要がある．

多重共線性を考慮した結果，説明変数として使用することとした特徴量を表2に示す．構文情報はg1-g36，単語心像性はg37-g38，文末表現はg39-g64，にそれぞれ対応している．

表2　既に採用している特徴量

g	特徴量
g1	助動詞(語彙数)
g2	接頭詞
g3	記号(語彙数)
g4	文数
g5	文の長さ平均(字数)
g6	カタカナ(語数)
g7	全角記号(語数)
g8	全角英数字(語数)
g9	形容詞(語数)
g10	副詞(語数)
g11	連体詞(語数)
g12	接続詞(語数)
g13	感動詞(語数)
g14	ひらがな(%)
g15	漢字(%)
g16	カタカナ(%)
g17	記号(%)
g18	TTR
g19	全角記号(%)
g20	英数字(%)
g21	全角英数字(%)
g22	名詞(%)
g23	形容詞(%)
g24	副詞(%)
g25	連体詞(%)
g26	接続詞(%)
g27	感動詞(%)
g28	「!」の数
g29	「?」の数
g30	句点の数
g31	読点の数
g32	中点の数
g33	3点リーダの数
g34	鍵括弧の数
g35	括弧の数
g36	「/」の数
g37	単語心像性4点台(語数)
g38	単語心像性6.5以上7.0未満(語数)
g39	か(語数)
g40	な(語数)
g41	し(語数)
g42	たい(語数)
g43	ない(語数)
g44	だ(文末語数)
g45	か(文末語数)
g46	な(文末語数)
g47	し(文末語数)
g48	です(文末語数)
g49	ます(文末語数)
g50	たい(文末語数)
g51	ない(文末語数)
g52	ぞ(%)
g53	だ(%)
g54	よ(%)
g55	ね(%)
g56	か(%)
g57	です(%)
g58	ます(%)
g59	ない(%)
g60	か(文末%)
g61	ですか(語数)
g62	ないです(語数)
g63	ますか(語数)
g64	ました(語数)

3.3. 推定結果

9 因子の因子得点を目的変数とし，64 個の文章の特徴量を説明変数として，ステップワイズ選択法による重回帰分析[13]を行った．この時の重相関係数の値を表 3 の「単項」の列に示す．重相関係数は，値が 0.9 以上ならば良好で，0.7 以上ならばやや良好，0.7 未満ならば不良である，という推定精度を表す[14]．第 1 因子から第 6 因子までの 6 因子は推定精度がやや良好であり，残りの第 7 因子から第 9 因子までの 3 因子は推定精度が不良であるという結果が得られた．

また，説明変数の積である二次の項も考慮した上で重回帰分析を行った．二次の項同士の多重共線性を考慮した結果，説明変数の数は 218 個となった．単項の場合と同様に，9 因子の因子得点を目的変数，218 個の特徴量を説明変数として，重回帰分析を行った．この時の重相関係数の値を表 3 の「二次項」の列に示す．第 3 因子と第 6 因子以外の 7 因子に関しては，値が 0.9 以上なので，推定精度が良好であるといえる．第 3 因子と第 6 因子は値が 0.9 に及ばなかったが，それでも推定精度はやや良好といえる．

さらに，二次の項に関しては，推定の良好性を推定誤差により評価する．実験の因子得点とその推定値の平均誤差の絶対値を求め，表 4 に示す．全体の誤差平均は非常に小さく，どの因子も因子得点に近い推定値が得られているといえる．

表 3 重相関係数(特徴量 64 個)

因子	重相関係数	
	単項	二次項
第1因子(的確性)	0.832	1.000
第2因子(不快性)	0.774	0.947
第3因子(独創性)	0.744	0.877
第4因子(容易性)	0.728	0.908
第5因子(執拗性)	0.893	0.966
第6因子(曖昧性)	0.872	0.899
第7因子(感動性)	0.581	0.997
第8因子(努力性)	0.650	0.904
第9因子(感動性)	0.683	0.954

表 4 各因子の残差の絶対値

因子	残差の絶対値
第1因子(的確性)	0.0000420
第2因子(不快性)	0.114
第3因子(独創性)	0.164
第4因子(容易性)	0.124
第5因子(執拗性)	0.0874
第6因子(曖昧性)	0.147
第7因子(感動性)	0.0202
第8因子(努力性)	0.112
第9因子(熱烈性)	0.0858
残差平均	0.0949

4. 新特徴量 -単語親密度と表記妥当性-

NTT データベースシリーズ[15]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[16]．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．既に特徴量として採用した単語心像性が，このようなデータに該当する．

ここでは，単語親密度と表記妥当性を新たに追加する．単語親密度とは，ある単語がどの程度なじみがあると感じられるかを 7 段階で表した指標である[17]．また，表記妥当性とは，ある単語の表記のもっともらしさを 5 段階で示した指標である[17]．例えば，乾電池の「たんに」という言葉例とすると，「単に」を意味する場合は単語親密度の値が 5.312，「乾電池」を意味する場合は値が 3.594 となる．これらの値は，被験者の得点を平均して求められている．

また，同じ単語の表記でも，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している．例えば，「娯楽」という単語を例とすると，ひらがな表記の場合は表記妥当性の値が 2.95，カタカナ表記の場合は 1.95，漢字表記の場合は 4.90，である．

このようにして，単語親密度，ならびに，表記妥当性の特徴量を抽出した．これらを表 5 に示す．特徴量としては，単語心像性の特徴量[3]と同様に，単語親密度，または，表記妥当性に該当した単語の数や該当した単語の割合，単語心像性の値が 1 点台(1.0～2.0 未満)，2 点台(2.0～3.0 未満)…のように，1 点間隔で特徴量をとったもの，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抽出した．表 5 に示した特徴量に対しても，多重共線性を考慮して説明変数を選出した．その結果，表 5 に網掛けを施した 13 個の特徴量を使用することとした．これらの特徴量を g65-g77 と称した上で，表 6 にまとめて示す．

表 5 単語親密度・表記妥当性の特徴量

単語親密度の該当単語(語彙数)
単語親密度の該当単語(語数)
単語親密度の該当単語率(語数)
単語親密度4点台(語彙数)
単語親密度4.5～5.0未満(語彙数)
単語親密度5点台(語彙数)
単語親密度5.0～5.5未満(語彙数)
単語親密度5.5～6.0未満(語彙数)
単語親密度6点台(語彙数)
単語親密度6.0～6.5未満(語彙数)
単語親密度6.5～7.0未満(語彙数)
単語親密度4点台(語数)
単語親密度4.5～5.0未満(語数)
単語親密度5点台(語数)
単語親密度5.0～5.5未満(語数)
単語親密度5.5～6.0未満(語数)
単語親密度6点台(語数)
単語親密度6.0～6.5未満(語数)
単語親密度6.5～7.0未満(語数)
表記妥当性の該当単語(語彙数)
表記妥当性の該当単語(語数)
表記妥当性の該当単語率(語数)
表記妥当性2点台(語彙数)
表記妥当性2.5～3.0未満(語彙数)
表記妥当性3点台(語彙数)
表記妥当性3.0～3.5未満(語彙数)
表記妥当性3.5～4.0未満(語彙数)
表記妥当性4点台(語彙数)
表記妥当性4.0～4.5未満(語彙数)
表記妥当性4.5～5.0未満(語彙数)
表記妥当性5点台(語彙数)
表記妥当性2点台(語数)
表記妥当性2.5～3.0未満(語数)
表記妥当性3点台(語数)
表記妥当性3.0～3.5未満(語数)
表記妥当性3.5～4.0未満(語数)
表記妥当性4点台(語数)
表記妥当性4.0～4.5未満(語数)
表記妥当性4.5～5.0未満(語数)
表記妥当性5点台(語数)

表 6 新たに使用する特徴量

g65	単語親密度の該当単語率(語数)
g66	単語親密度6.5～7.0未満(語彙数)
g67	単語親密度4点台(語数)
g68	単語親密度5点台(語数)
g69	単語親密度5.5～6.0未満(語数)
g70	単語親密度6点台(語数)
g71	単語親密度6.0～6.5未満(語数)
g72	表記妥当性の該当単語率(語数)
g73	表記妥当性3点台(語数)
g74	表記妥当性3.5～4.0未満(語数)
g75	表記妥当性4点台(語数)
g76	表記妥当性4.0～4.5未満(語数)
g77	表記妥当性5点台(語数)

5. 因子得点の推定結果

5.1. 単項のみを考慮した重回帰分析

印象評価実験で使用した 60 個の質問回答文に対し，表 2 と表 6 に示す 77 個の特徴量を説明変数とし，表 1 に示す 9 因子の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[13]を行った．分析の結果，重回帰式(1)が得られた．

$$
\left.
\begin{aligned}
y_1 &= 0.00295g_{37} + 0.150g_{12} + 0.0609g_{59} + 0.0533g_{21} \\
&\quad + 0.0105g_7 - 0.113g_{29} + 0.0896g_9 + 0.0151g_{19} \\
&\quad - 0.750 \\
y_2 &= 1.73g_{44} + 1.79g_{50} - 0.115g_1 + 0.110g_{43} + 0.0341g_8 \\
&\quad - 0.172g_{12} + 0.261 \\
y_3 &= 0.161g_{52} + 0.0682g_8 - 0.0670g_{43} + 0.888g_{50} \\
&\quad + 0.0816g_{60} - 0.0852g_{74} + 0.0197g_{65} - 0.175 \\
y_4 &= -0.0468g_{37} - 0.0843g_{52} + 0.0916g_9 - 0.178g_{45} \\
&\quad + 0.305 \\
y_5 &= 0.0807g_1 + 0.0168g_6 + 0.126g_{29} + 0.0155g_{15} \\
&\quad + 0.00308g_5 + 0.0275g_4 + 0.502g_{49} + 0.731g_{50} \\
&\quad - 0.0792g_{70} - 1.14 \\
y_6 &= -0.0759g_{31} - 0.11163g_{43} - 0.0618g_{56} - 0.0342g_8 \\
&\quad - 0.126g_{55} - 0.772g_{44} - 0.691g_{50} - 0.125g_{25} + 0.698 \\
y_7 &= 0.557g_{13} + 0.0297g_{66} + 0.105g_{23} - 0.243 \\
y_8 &= 0.0871g_1 + 0.0788g_{60} + 0.104g_{23} - 0.00560g_{20} \\
&\quad - 0.00905g_8 - 0.388 \\
y_9 &= 0.193g_{52} - 0.0435g_{75} - 0.0148g_{16} + 0.321g_{50} \\
&\quad - 0.109g_{51} - 0.0260
\end{aligned}
\right\} \quad (1)
$$

この時の重相関係数を表 7 の「単項」の列に示す．第 1 因子から第 6 因子までの 6 因子中，第 5 因子のみは推定精度が良好で，残りの 5 因子は推定精度がやや良好であるといえる結果になった．また，第 7 因子から第 9 因子までの 3 因子に関しては，推定精度が不良であるという結果が得られた．

5.2. 二次の項を考慮した重回帰分析

次に，単項のみの分析で使用した 77 個の説明変数に関して，二次の項を考慮する．二次の項同士の多重共線性を考慮した結果，説明変数の数は 281 個となった．単項の場合と同様に，印象評価実験で使用した

$$
\begin{aligned}
y_1 &= 0.210g_{12}g_{58} + 0.0369g_{50}g_{60} + 0.174g_{30}g_{66} - 0.0909g_{12}g_{6-} - 0.413g_{54}g_{62} - 0.135g_{40}g_{66} + 0.102g_{13}g_{15} + 0.166g_{24}g_{61} - 0.180g_{2<}g_{-1} \\
&+ 0.704g_{51}g_{60} - 0.0893g_{13}g_{<-} - 0.0535g_{52}g_{-3} + 0.149g_{42}g_{<<} + 0.000737g_{5}g_{<-} - 0.0904g_{26}g_{55} + 0.0527g_{32}g_{42} - 0.0504g_{35}g_{18} \\
&- 0.0781g_{21}g_{60} + 0.0142g_{1<}g_{<1} + 0.0174g_{21}g_{42} - 0.00404g_{14}g_{18} - 0.338
\end{aligned}
$$

$$
\begin{aligned}
y_2 &= -0.0486g_{40}g_{5-} + 0.00356g_{-2}g_{-<} - 0.0447g_{32}g_{42} - 0.0111g_{2<}g_{-2} - 0.0141g_{60}g_{-2} - 0.176g_{13}g_{--} - 0.0287g_{12}g_{16} + 1.01g_{28}g_{<3} \\
&+ 0.151g_{20}g_{<<} + 0.165g_{60}g_{-6} - 0.188g_{21}g_{60} - 0.0509g_{49}g_{54} + 0.0415g_{1}g_{18} - 0.325g_{46}g_{69} + 0.0479g_{8}g_{<0} + 0.0345g_{19}g_{51} \\
&+ 0.0288g_{-1}g_{-3} - 0.0568g_{51}g_{60} + 0.00155g_{21}g_{22} - 0.0512g_{43}g_{54} + 0.140g_{32}g_{61} - 0.0196g_{30}g_{66} - 0.00425g_{16}g_{5-} + 0.00648g_{9}g_{16} \\
&- 0.103g_{13}g_{-3} - 0.0921g_{38}g_{18} - 0.504g_{60}g_{64} + 0.0120g_{30}g_{56} + 0.000152g_{<5}g_{16} - 0.00898g_{8}g_{39} - 0.0185g_{21}g_{60} + 0.0412g_{60}g_{-1} \\
&+ 0.00388g_{6}g_{5-} + 0.0204g_{-6}g_{--} - 0.0562g_{55}g_{-6} + 0.00349g_{43}g_{49} + 0.00213g_{4}g_{58} + 0.000243g_{<}g_{24} + 0.00106g_{9}g_{-2} - 0.00116g_{-4}g_{-<} \\
&+ 0.00458g_{6}g_{54} - 0.0000604g_{22}g_{-2} - 0.000833g_{22}g_{-2} - 0.000833g_{-}g_{-5} - 0.00486g_{49}g_{55} + 0.00583g_{25}g_{32} - 0.00219g_{32}g_{54} \\
&+ 0.000484g_{54}g_{5-} - 0.000809g_{41}g_{-6} + 0.00319g_{64}g_{-4} + 0.00120g_{-1}g_{<4} - 0.000160g_{51}g_{-2} + 0.0000497g_{16}g_{48} - 0.00000922g_{3}g_{22} \\
&- 0.0000527g_{25}g_{-5} + 0.00124g_{2-}g_{-<} + 0.00000238g_{29}g_{-3} - 0.00000274g_{<5}g_{-3} + 0.000000177g_{10}g_{<0} - 0.0591
\end{aligned}
$$

$$
\begin{aligned}
y_3 &= 0.0179g_{52}g_{-3} + 0.979g_{22}g_{<3} + 0.113g_{60}g_{-1} - 0.00832g_{33}g_{59} + 0.298g_{2<}g_{60} + 0.0136g_{5<}g_{-2} + 0.0300g_{6<}g_{77} - 0.0212g_{5}g_{5-} \\
&- 0.425g_{<4}g_{63} + 0.0192g_{8}g_{39} - 0.0100g_{16}g_{60} - 0.106g_{<<}g_{-3} - 0.0153g_{19}g_{-4} + 0.176g_{40}g_{<4} + 0.0606g_{24}g_{40} - 0.00110g_{20}g_{58} \\
&- 0.0150g_{49}g_{5-} + 0.450g_{51}g_{60} - 0.0310g_{32}g_{42} + 0.214g_{46}g_{69} - 0.824g_{-}g_{-6} + 0.404g_{35}g_{45} - 0.0331g_{11}g_{50} - 0.128g_{40}g_{<<} \\
&+ 0.0387g_{1}g_{18} + 0.0514g_{2-}g_{36} - 0.111g_{4<}g_{-4} + 0.0386g_{60}g_{--} - 0.227g_{4<}g_{53} + 0.0120g_{0}g_{<-} - 0.101g_{63}g_{-1} - 0.0205g_{24}g_{54} \\
&- 0.00159g_{16}g_{5-} + 0.00128g_{16}g_{-5} - 0.00672g_{12}g_{58} - 0.00299g_{66}g_{-2} - 0.00302g_{56}g_{<0} + 0.0000512g_{5}g_{16} - 0.00122g_{30}g_{65} - 0.0856
\end{aligned}
$$

$$
\begin{aligned}
y_4 &= -0.178g_{48}g_{--} + 0.00717g_{1<}g_{<-} - 0.000238g_{55}g_{6} - 0.486g_{33}g_{4<} - 0.0474g_{46}g_{5-} + 0.961g_{33}g_{-6} - 0.0612g_{<2}g_{-1} - 0.0119g_{-9}g_{-2} \\
&- 0.0734g_{41}g_{<0} + 0.00156g_{3}g_{22} - 0.00339g_{16}g_{3-} + 0.553g_{<5}g_{-1} + 0.00485g_{31}g_{54} + 0.0284g_{10}g_{65} + 0.0116g_{-2}g_{6} + 0.466g_{51}g_{60} \\
&+ 0.0855g_{2<}g_{<0} - 0.106g_{54}g_{64} - 0.0141g_{<6}g_{65} - 0.0232g_{10}g_{2<} - 0.0173g_{51}g_{-2} + 0.00435g_{4}g_{46} - 0.219g_{60}g_{66} - 0.0639g_{2}g_{56} \\
&- 0.0105g_{-}g_{-5} + 0.0334g_{25}g_{32} + 0.0406g_{1-}g_{<9} - 0.0248g_{49}g_{59} + 0.109g_{60}g_{69} + 0.00289g_{5<}g_{-2} + 0.376g_{25}g_{66} + 0.0421g_{18}g_{58} \\
&- 0.0490g_{49}g_{66} + 0.0163g_{53}g_{--} - 0.124g_{46}g_{-0} - 0.000393g_{20}g_{<8} + 0.00546g_{46}g_{56} + 0.00258g_{19}g_{24} + 0.000827g_{68}g_{-2} \\
&- 0.00177g_{16}g_{-0} + 0.00365g_{71}g_{-3} + 0.00925g_{61}g_{-5} - 0.00218g_{19}g_{<<} - 0.000573g_{9}g_{<6} + 0.0145g_{28}g_{53} - 0.0405g_{2<}g_{35} \\
&- 0.00269g_{32}g_{61} + 0.00531g_{2<}g_{-1} + 0.000473g_{6<}g_{-6} + 0.000959g_{<5}g_{46} + 0.000713g_{2-}g_{36} - 0.0000444g_{22}g_{5-} - 0.000255g_{56}g_{68} \\
&- 0.0000329g_{56}g_{<0} - 0.0000544g_{-2}g_{--} - 0.000268g_{63}g_{68} - 0.00000214g_{4<}g_{<0} + 0.0000000382g_{15}g_{18} + 0.183
\end{aligned}
$$

$$
\begin{aligned}
y_5 &= 0.182g_{1}g_{18} + 0.000280g_{5}g_{6} + 0.00467g_{24}g_{60} - 0.0467g_{23}g_{<8} + 0.00985g_{4}g_{<8} + 0.102g_{23}g_{29} + 0.339g_{33}g_{4<} - 0.201g_{66}g_{-1} \\
&- 0.0149g_{51}g_{-2} - 0.266g_{<1}g_{54} - 0.672
\end{aligned}
$$

$$
\begin{aligned}
y_6 &= -0.169g_{1}g_{18} - 0.0548g_{31}g_{44} - 0.0207g_{22}g_{<3} + 0.0826g_{49}g_{<4} - 0.00124g_{5}g_{22} - 0.0144g_{<5}g_{-2} - 0.109g_{60}g_{77} - 0.0204g_{4}g_{18} \\
&- 0.0739g_{68}g_{--} - 0.0749g_{9}g_{54} + 0.0247g_{32}g_{60} + 0.0190g_{68}g_{-2} + 0.00864g_{1<}g_{23} - 0.00686g_{16}g_{-0} + 0.110g_{48}g_{63} - 0.116g_{25}g_{-3} \\
&- 0.0563g_{24}g_{40} + 0.0202g_{69}g_{-<} - 0.0503g_{<2}g_{-3} - 0.0135g_{21}g_{<6} + 0.0504g_{54}g_{--} + 0.0181g_{16}g_{<1} - 0.332g_{51}g_{60} + 0.0901g_{19}g_{60} \\
&+ 0.105g_{3}g_{11} + 0.114g_{<4}g_{69} + 0.0200g_{29}g_{-1} - 0.00800g_{-}g_{2<} + 0.0375g_{25}g_{56} + 0.0396g_{2<}g_{-5} + 0.00247g_{16}g_{<-} - 0.403g_{2<}g_{66} \\
&- 0.00561g_{16}g_{48} + 0.00799g_{<4}g_{5<} + 0.129g_{63}g_{-1} - 0.000959g_{<}g_{<4} - 0.0107g_{-3}g_{-3} - 0.00945g_{45}g_{<9} - 0.00362g_{28}g_{6<} \\
&- 0.00355g_{25}g_{58} - 0.0119g_{2}g_{<6} + 0.000505g_{5}g_{-2} + 0.00146g_{16}g_{66} - 0.00135g_{32}g_{59} - 0.00360g_{29}g_{54} - 0.00124g_{49}g_{56} \\
&+ 0.0000572g_{14}g_{18} + 0.00960g_{4-}g_{69} + 0.00000334g_{<}g_{5-} + 0.000212g_{-6}g_{--} + 0.0000234g_{1-}g_{58} + 0.00000263g_{<}g_{65} \\
&- 0.0000250g_{60}g_{-5} + 0.0000166g_{12}g_{-0} - 0.00000330g_{0}g_{55} + 0.000000122g_{20}g_{-3} - 0.0000000433g_{22}g_{59} \\
&+ 0.000000101g_{-4}g_{--} + 0.795
\end{aligned}
$$

$$
\begin{aligned}
y_7 &= 0.0258g_{10}g_{6<} + 0.643g_{<1}g_{61} + 0.143g_{13}g_{<-} + 0.106g_{30}g_{66} - 0.0235g_{29}g_{-3} + 0.00282g_{<}g_{64} + 0.0296g_{28}g_{6<} + 0.459g_{28}g_{<3} \\
&+ 0.0857g_{52}g_{-3} - 1.28g_{60}g_{61} + 0.708g_{51}g_{60} + 0.0226g_{23}g_{58} + 0.0988g_{45}g_{62} + 0.0351g_{12}g_{50} - 0.00278g_{21}g_{22} + 0.0190g_{50}g_{54} \\
&- 0.00435g_{1-}g_{56} + 0.0394g_{26}g_{55} - 0.326
\end{aligned}
$$

$$
\begin{aligned}
y_8 &= 0.274g_{23}g_{42} + 0.834g_{13}g_{-3} + 0.102g_{11}g_{68} + 0.0571g_{60}g_{--} + 0.0457g_{40}g_{5-} + 0.300g_{60}g_{69} - 0.0471g_{32}g_{58} - 0.223g_{66}g_{-1} \\
&+ 0.00873g_{16}g_{48} - 0.253g_{26}g_{<4} + 0.00512g_{4}g_{56} - 0.0422g_{29}g_{-1} - 0.299g_{30}g_{66} + 0.0529g_{32}g_{40} + 0.143g_{60}g_{66} - 0.0585g_{61}g_{-<} \\
&- 0.00436g_{10}g_{24} - 0.232
\end{aligned}
$$

$$
\begin{aligned}
y_9 &= 0.133g_{42}g_{<0} + 0.103g_{<2}g_{-3} + 0.192g_{16}g_{-6} + 0.0242g_{<6}g_{65} + 0.0238g_{-4}g_{-<} - 0.310g_{4-}g_{<4} - 0.0532g_{56}g_{<0} - 0.000204g_{5}g_{0} \\
&+ 0.0422g_{54}g_{<5} + 0.0313g_{9}g_{<0} - 0.141g_{18}g_{5-} - 0.113g_{32}g_{42} + 0.254g_{42}g_{--} - 0.0501g_{60}g_{-6} + 0.0372g_{23}g_{<-} + 0.514g_{25}g_{66} \\
&+ 0.0442g_{2-}g_{36} + 0.319g_{60}g_{64} - 0.0290g_{8}g_{<0} - 0.00793g_{4}g_{18} - 0.00400g_{16}g_{23} - 0.00750g_{19}g_{25} + 0.0202g_{26}g_{65} \\
&- 0.204g_{55}g_{-6} + 0.0163g_{16}g_{51} + 0.0202g_{54}g_{-3} - 0.0427g_{2<}g_{48} - 0.0987g_{48}g_{63} - 0.00774g_{2<}g_{-2} + 0.00371g_{29}g_{72} \\
&+ 0.0383g_{12}g_{61} - 0.0251g_{49}g_{60} + 0.00784g_{11}g_{<8} + 0.0223g_{2<}g_{10} + 0.0389g_{2-}g_{-<} - 0.0142g_{1<}g_{-3} - 0.0125g_{25}g_{-5} \\
&+ 0.0157g_{43}g_{-1} + 0.00123g_{21}g_{42} - 0.0350g_{51}g_{69} - 0.000426g_{16}g_{-<} - 0.00543g_{11}g_{6<} + 0.0716
\end{aligned}
$$

(2)

60 個の質問回答文に対し，281 個の特徴量を説明変数とし，9 因子の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[13]を行った．この結果，重回帰式(2)が得られた．この時の重相関係数を表 7 の「二次項」の列に示す．どの因子も重相関係数の値が 0.9 以上となっており，どの因子も推定精度が良好であるといえる．

6. 考察

3.で示した文章の特徴量が64個の場合と，5.で示した文章の特徴量が77個の場合とで，重相関係数の値が変動しているかどうかを比較検討する．表3と表7の結果を表8にまとめて表す．単項のみの場合，第3因子と第5因子に関しては値が向上しているが，第7因子と第9因子に関しては，値がやや低下している．残りの5因子に関しては，値に変動が見られなかった．

二次項に関しては，第2因子，第3因子，第4因子，第6因子，第8因子，第9因子の6因子に関しては値が向上している．一方で，第 1 因子，第 5 因子，第 7 因子の 3 因子に関しては値が低下している．しかしながら，特徴量が 64 個の場合では重相関係数が 0.9 に及ばなかった第3因子と第6因子に関しては，値が大幅に向上している．結果的には，全ての因子において重相関係数の値が 0.9 を上回っているため，どの因子も推定精度が良好であるといえる結果となっている．従って，文章の特徴量を追加したことにより，推定精度の更なる向上に成功したといえる．

表 7 重相関係数(特徴量 77 個)

因子	重相関係数	
	単項	二次項
第1因子(的確性)	0.832	0.989
第2因子(不快性)	0.774	1.000
第3因子(独創性)	0.788	0.999
第4因子(容易性)	0.728	1.000
第5因子(執拗性)	0.908	0.925
第6因子(曖昧性)	0.872	1.000
第7因子(感動性)	0.552	0.963
第8因子(努力性)	0.650	0.950
第9因子(感動性)	0.674	1.000

表 8 重相関係数の比較

因子	特徴量64個の場合		特徴量77個の場合	
	単項	二次項	単項	二次項
第1因子(的確性)	0.832	1.000	0.832	0.989
第2因子(不快性)	0.774	0.947	0.774	1.000
第3因子(独創性)	0.744	0.877	0.788	0.999
第4因子(容易性)	0.728	0.908	0.728	1.000
第5因子(執拗性)	0.893	0.966	0.908	0.925
第6因子(曖昧性)	0.872	0.899	0.872	1.000
第7因子(感動性)	0.581	0.997	0.552	0.963
第8因子(努力性)	0.650	0.904	0.650	0.950
第9因子(感動性)	0.683	0.954	0.674	1.000

7. まとめ

本論文では，質問者と回答者の相性の判定をめざして，文章の因子得点の推定精度の向上を行った．ここでは，構文情報，単語心像性と文末表現に加え，単語親密度と表記妥当性を文章の特徴量として使用し，質問回答文の因子得点の推定を試みた．その結果，全因子において精度良く推定できることを示した．

今後は，質問文に対して適切な回答が見込める利用者を選出する手法の確立を行い，そのようなプロトタイプシステムの構築・評価を行う予定である．また，今回得られた重回帰式を用いて，まだ求まっていない質問回答文の因子得点を求める．これをもとに「ベストアンサー」の推定を行い，その推定手法を確立させていく．更に，客観的な視野が加味された「グッドアンサー」[18]の推定手法も確立した上で，「ベストアンサー」との比較検討を行っていく予定である．

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] Yahoo!知恵袋, http://chiebukuro.yahoo.co.jp/

[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印象の因子得点の推定，第 14 回日本感性工学会大会，B1-01, 2012.

[4] Blooma, M.J. and Chua, A.Y.K. and Goh, D.H.L.: A Predictive Framework for Retrieving the Best Answer, Proc. of 2008 ACM Symposium on Applied Computing (SAC08), pp.1107-1111, 2008.

[5] Agichtein, E., Castillo, C., Donato, D., Gionis, A. and Mishne, G.:Finding High-Quality Content in Social Media, Proc. of the Int'l Conf. on Web Search and Web Data Mining (WSDM08), pp.183-194, 2008.

[6] Wang, X. J., Tu, X., Feng, D. and Zhang, L.: Ranking Community Answers by Modeling Question-Answer Relationships via Analogical

Reasoning, Proc. of 32nd Int'l ACM SIGIR Conf., pp.179-186, 2009.

[7] Kim, S., Oh, J. S. and Oh, S.:Best-Answer Selection Criteria in a Social Q&A site from the User-Oriented Relevance Perspective, Proc. of American Society for Information Science and Technology (ASIS&T) 2007 Annual Meeting, 2007.

[8] Adamic, L. A., Zhang, J., Bakshy, E. and Ackerman, M. S.:Knowledge Sharing and Yahoo Answers: Everyone Knows Something, Proc. of 17th Int'l Conf. on World Wide Web (WWW2008), 2008.

[9] Jurczyk, P. and Agichtein, E.: Discovering Authorities in Question Answer Communities by Using Link Analysis, Proc. of 16th ACM Conf. on Inf. and Know. Management (CIKM2007), pp.919-922, 2007.

[10] Hovy, E., Gerber, L., Hermjakob, U., Junk, M. and Lin, C.-Y.: Question Answering in Webclopedia, Proc. of 9th Text Retrieval Conf., pp. 655-664, 2000.

[11] 西原陽子, 松村真宏, 谷内田正彦:Q&A コミュニティでの質疑応答パターンの理解, 第 22 回人工知能学会全国大会, 1H2-7, 2008.

[12] 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.

[13] 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

[14] 菅民郎:初心者がらくらく読める多変量解析の実践上, (社), p.39, (社)現代数学社, 1993.

[15] 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.

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

[17] 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 1 期, (社)三省堂, 2003.

[18] 情報学研究データリポジトリ :「Yahoo!知恵袋データ(第 2 版)」の提供について, http://www.nii.ac.jp/cscenter/idr/yahoo/chiebkr2/Y_chiebukuro.html