社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－NL－164（8）
2004／11／4

フェイスシートとの関係を利用した自由回答アンケートの分析

内山　将夫†　　大塚　裕子††　　井佐原　均†

† 情報通信研究機構　〒619-0289 京都府相楽郡精華町光台 3-5
†† 計量計画研究所　〒162-0845　東京都新宿区市ヶ谷本村町 2 番 9 号
E-mail: †{mutiyama,isahara}@nict.go.jp, ††hinui@ibs.or.jp

あらまし　本稿では，自由回答アンケートに関わる確率的な構造を示すとともに，そこから，情報理論，とくに，相互情報量を利用することにより，自由回答アンケートの分析手法が統一的に導出可能なことを示す．
キーワード　自由回答アンケート，分析，情報理論，相互情報量

Open-ended Questionnaire Analysis based on the Relationship with the Face Sheet

Masao UTIYAMA†, Hiroko OTSUKA††, and Hitoshi ISAHARA†

† National Institute of Information and Communications Technology　Seika-cho, Soraku-gun, Kyoto, Japan
†† The Institute of Behavioral Sciences　2-9 Motomura-cho, Ichigaya, Shinjuku-ku, Tokyo 162-0845 Japan
E-mail: †{mutiyama,isahara}@nict.go.jp, ††hinui@ibs.or.jp

Abstract　This paper proposes a probabilistic structure of open-ended questionnaires (OEQ) and derives a set of information-theoretic analysis methods for OEQ.
Key words　open-ended questionnaire analysis, information theory, mutual information

1. はじめに

われわれは，自由回答アンケートの回答テキストから調査者および回答者に必要な情報の抽出および分類を行うことを目的に，これらの情報を，1）どのような観点で分類すべきか，2）必要な観点で情報を取り出すために有効な手がかりは何か，を明らかにしようとしてきた[1]．これまでの研究を踏まえ，本稿ではさらに自由回答テキストの多面的な処理の研究を深めるために，アンケート票のフェイスシートを利用した客観的な自由回答アンケート分析を提案する．このような研究では，処理対象であるアンケート回答のテキストの性質について十分に考慮する必要がある．

アンケート調査は，調査の実施までに下記のプロセスをたどる．

（1）　問題意識を持つ
（2）　調査目的を明確にする
（3）　仮説を設定し，調査項目を列挙する
（4）　質問項目に絞り込む
（5）　質問に変換する

上記に示すように，アンケート中の質問項目は調査目的を基盤に作成されることが重要である．質問項目を「調査対象のどのような側面を調べようとするのか」という観点から分類すると，表1のようになる．この表で，1と2を客観的に把握することが可能な質問項目を，とくにフェイス項目とよぶ．フェイス項目とよぶのは，これらの質問項目が，調査票の顔（フェイス）に該当する第1ページ目におかれることが従来のアンケートでは多かったからである[2]．

しかし，質問項目が上述の調査実施までのプロセスに示したとおり調査目的および項目に関連したものであるのに対し，フェイス項目はそれ自体が調査対象になる項目ではない．フェイス項目は，次の二つの理由，1）なぜ回答内容が異なるのかを説明する要因，ならびに，2）回答者に偏りがないかどうかを確認する基準という役割を持っている．1）は，たとえば「横浜北西線道路計画に関する多様な意見を知る」という調査目的に対し，回答者の回答内容に違いがあるのはなぜかを知る手がかりとして設定する．具体的には「横浜北西線道路計画に関する多様な意見」のように，回答内容が異なる要因を知りたい質問項目を目的変数（従属変数），目的変数の違いを説明するための質問項目を説明変数（独立変数）として結果を解析する．したがって，フェイス項目は，目的変数によって，得られた自由回答を解析するうえで重要な分析観点の一つである．このような目的のために，調査目的とは無関係のフェイス項目がアンケート調査の質問項目に加えられる．

フェイス項目には，従来の国勢調査で問うような世帯人数，

表 1　質問項目の分類（文献 [2] より引用）

質問項目の内容	個人を対象とする調査での質問項目の例	企業を対象とする調査での質問項目の例
1. 調査対象の基本的属性	性別，年齢，学歴，職業，年収	業種，資本金規模，従業員数
2. 調査対象の所属する集団の属性	世帯員数，世帯構成，居住地域	立地地域，加盟している業界・団体
3. 調査対象の所有物	自家用自動車の有無	ＯＡ機器導入状況
4. 調査対象の活動や行動	旅行先，購入商品名	新卒者雇用実績，取り扱い商品数
5. 調査対象の意志や態度	住宅購入予定，いじめの対応	新卒者採用予定数
6. 調査対象のもつ知識	新製品の知名	新規立地特別減税制度の存在の知名
7. 調査対象のもつ経験や習慣	図書館の利用回数，ふだん見るテレビ番組	取引実績
8. 調査対象のもつ意見や願望	地域社会の活性策，学制改革案に対する賛否	産業政策への要望
9. 調査対象の行動の理由や動機	商品購入理由，大学入学動機	工場進出理由
10. 調査対象の感覚や印象	製品の使用感，商店街のイメージ	景気の先行き感

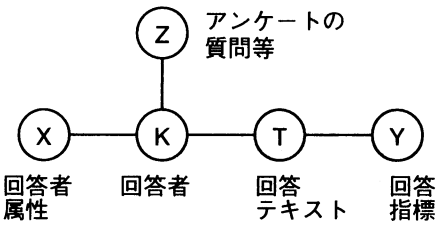

図 1　自由回答アンケートにおけるマルコフ性
Fig. 1　Markov field of the open-ended questionnaire

世帯主との続柄，年齢，性別，職業など人口統計学的な分類基準に加え，対象者の感性や知性についての特性，どんな感じ方，考え方，関心をもち，どんな性格やライフスタイルかといった特性（心理的特性），調査テーマと直接かかわりあるものごとについての経験を示す特性（経験的特性），来場者調査や交通に関するアンケートの場合は，居住地，最寄り駅（地理的特性），その他（特殊な特性）などがある [3]．しかし，フェイス項目は，例えば収入および家計支出のように，生活水準・生活態度を示す指標として重要であるが，プライバシーにかかわるので応答拒否が多いものもある [4]．したがって，応答拒否を避けるために，フェイスシートをすべての質問の最後に添付したり，本研究で扱う横浜北西線の道路計画によせられる意見(注1)のように，予めフェイス項目を居住地だけに限定したりする場合も出てきている．フェイス項目がどの程度，収集された回答の分析において影響を与えるのかを知ることは重要な課題である．

この課題を達成するために，我々は，情報理論 [5] を利用した分析を提案する．本稿では，まず，自由回答アンケートの情報構造を示し，そこから，相互情報量を利用することにより，各種分析手法が導出可能なことを示す．

2. 情報構造からの分析方法の導出

自由回答アンケートに関わる要素である確率変数間には，図 1 のマルコフ性があることを我々は想定した．図 1 において，X，K，T，Y は，それぞれ，回答者属性，回答者，回答テキスト，回答指標であり，Z は，回答者に与えられたアンケートの質問等である．図 1 において，実線で結ばれている変数間は確率的に依存しているが，そうでない変数間は，条件付きで独立である．

ここで，回答者属性 X とは，「居住区」等のフェイス項目における回答者の属性であり，回答者 K とは，自由回答アンケートに対する回答者であり，回答テキスト T とは回答者が書いた自由回答アンケートのテキストであり，回答指標 Y とは，後述するように，回答テキストから抽出可能な情報であり，複数の観点からのアンケート分析のために利用する．また，Z は，回答者に与えられた自由回答アンケートにおける質問を代表とするような，回答者に対する，その他一切の入力情報である．

図 1 の要素のうち，本稿においては，回答者 K と入力 Z は，分析に利用しない(注2)．そのため，これらを除外した

$$X \rightarrow T \rightarrow Y \tag{1}$$

というマルコフ連鎖に基づき，分析方法を導出する．

さて，図 1 や (1) 式において，回答者属性 X と回答テキスト T は，分析対象としては所与である．そもそも，我々の目標は，これら 2 つの関係を分析することである．その分析の観点を，回答指標である Y が提供する．

回答指標 Y は，回答テキストのみから抽出可能でなければならないという制約があるが，その指標の選択は，分析の観点に応じたものを選べば良いという自由度がある．この分析の観点を如何なるものにするかは，分析対象とするアンケートに依存するものであるため，この観点の設定自体が重要な研究対象である．たとえば，本稿の実験においては，道路建設に関わるパブリック・インボルブメント (PI) における自由回答テキストを分析対象とするため，将来的には，Y の指標値として，{環境保護に関する回答, 道路建設の推進に関する回答，無関心,...} のように，各回答が述べている内容を，指標値として適切に抽出し，それら指標値と X の属性値との関係を求めたい．たとえば，回答属性 X が「居住区」であるとき，その属性値である「A 区」には「道路建設の推進に関する回答」が多いとか，「B 区」には「環境保護に関する回答」が多いとか，を求めたい．

しかし，我々は，現状においては，そのような PI の回答分析に必要な観点自体を模索している段階である．そのため，本稿においては，特別な分析の観点は設定せずに，後述するよう

(注1)：http://www.yokohama-nwline.jp/ref_pi/

(注2)：図 1 において，回答者 K は，概念的に設定したものであり，それを実際に利用することは想定していない．一方，回答者に対する入力 Z については，利用することは原理的には可能であるが，本稿においては利用していない．なお，Z の利用が必要なことは後述する．

に,(1) 回答テキスト中の単語と (2) 回答クラスタ (クラスタリング方法は後述) を，回答指標 Y として利用する．これらの指標を利用すると,「A区」の回答には「便利，ルート，インターチェンジ」という単語が多いとか,「回答クラスタ1，2，3」が多いとかが分かる．

このような X と Y との関係は，情報理論における通信路のアナロジーにより，解釈できる．つまり，(1) 式では，入力である属性 X が，ノイズのある通信路 T を通ることにより，指標 Y として出力されると考える．このアナロジーから，入力 X が出力 Y に，どれだけ関係しているかを，X と Y との相互情報量 $I(X;Y)$ で表すことが妥当だと言える．$I(X;Y)$ が大きいときには，X から Y，あるいは，Y から X が予測できる．たとえば，上例では,「道路建設の推進に関する回答」が多い区は「A区」であるとかが分かる．すなわち，$I(X;Y)$ が大きいときには，属性 X と指標 Y とは，密接に関係する．

したがって，相互情報量を利用することにより，同一の指標 Y について，2つの属性 X と X' のどちらの影響がより大きいかを，$I(X;Y)$ と $I(X';Y)$ とを比較することにより，推測できる．たとえば，指標 Y として道路建設への「賛否」としての {賛成，反対，その他} という指標値があるとき，X が「居住区」で，X' が「性別」のとき，I(居住区;賛否) と I(性別;賛否) とを比較し，$I(\cdot)$ が大きい属性の方が，指標との関係も強いと判断できる．なお，自然言語処理においては，2単語 X と Y の共起の強さを測定する尺度として，$I(\cdot)$ と等価な対数尤度比が有効である [6] ことが知られている．

次に，属性 X と指標 Y とが固定されたときに，属性値の重要性を測定する方法を示す．属性 X の属性値を $\{x_1, x_2, \ldots\}$ とし，指標 Y の指標値を $\{y_1, y_2, \ldots\}$ とする．このとき，相互情報量 $I(X;Y)$ は，以下のように式変形できる．

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x} p(x) \sum_{y} p(y|x) \log \frac{p(y|x)}{p(y)} \\ &= \sum_{x} p(x) D(p(y|x)||p(y)) \end{aligned} \tag{2}$$

ただし，$D(\cdot)$ はダイバージェンス [5] である．

(2) 式より，属性 X における，属性値 x の重要性として

$$\nu(x;Y) = p(x) D(p(y|x)||p(y)) \tag{3}$$

を利用することが考えられる．$\nu(x;Y)$ を属性値 x の寄与度と定義する．$\nu(\cdot)$ が大きい属性値ほど，相互情報量を大きくすることに貢献する．また，寄与率として $\tau(x;Y) = \frac{\nu(x;Y)}{I(X;Y)}$ を定義する．寄与率は0から1の値をとる．

この寄与度については，通信路のアナロジーから言うと，$\nu(x;Y)$ が大きい入力である属性値 x は，出力としての指標 Y に，大きく反映されると言える．そのことは (3) 式より，$\nu(x;Y)$ が大きいためには，まず，$p(x)$ が大きいことから，x は多くの回答に付随すること，次に，$D(\cdot)$ が大きいことから，x を条件とする Y の分布 $p(y|x)$ が全体での分布 $p(y)$ より離れていること，つまり，x を条件とする Y が特徴的な分布をしていることがいえる．このように，数が多く，かつ，特徴的な属性値は，そのまま指標にも反映することが言える．

たとえば，本稿の道路建設におけるアンケート回答においては,「地域」の属性において,「都筑区」,「その他の横浜市」... という順番で，寄与率の降順にソートできる．この上位2つの地域は，それぞれ，回答の度数も大きい．また,「都筑区」は「その他の横浜市」より度数は小さいが，寄与率は大きいので,「都筑区」の回答は,「その他の横浜市」よりも特色があると推測できる．そして，このことは,「都筑区」が道路建設により関わりの深い地域であることから，妥当な推測であると言える．

最後に指標値の重要性について述べる．(2) 式より，属性値 x に対する，指標値 y の重要性として

$$d(y|x) = p(y|x) \log \frac{p(y|x)}{p(y)} \tag{4}$$

を利用する．$d(y|x)$ が大きい指標値ほど，$I(X;Y)$ に貢献する．$d(y|x)$ が大きい指標値 y は，回答全体における確率 $p(y)$ よりも，その属性値 x における確率 $p(y|x)$ が大きいので，$d(y|x)$ は，属性値 x に対する指標値 y の特徴度と考えることができる．

たとえば，上例の「都筑区」において特徴度の高い単語は「都筑，便利，ルート，インターチェンジ，なる，よい，心配，港北，ほしい，ば，騒音，ニュータウン」である．これより,「ルートができれば便利になる反面，騒音などが心配である」というような意見があることが推測できる．このような推測は，特徴度の高い回答クラスタを読むことにより，妥当であることが確認できる．

なお，文献 [7] では，$\frac{d(y|x)}{D(p(y|x)||p(y))}$ と等価な尺度を，動詞 x に対する名詞カテゴリ y の共起の適切さの測定尺度として利用しているが，その導出の動機や方法は本稿とは異なる．

次に，本稿で利用した回答指標 Y について述べる．

3. 回答指標

(2) 式の計算に必要な確率 $p(x,y)$ の推定は，本稿では，属性値 x と指標値 y との共起頻度 $f(x,y)$ により，以下で推定した．

$$p(x,y) = \frac{\mathrm{freq}(x,y)}{\sum_{x,y} \mathrm{freq}(x,y)} \tag{5}$$

このようにしたとき，単語(注3)を回答指標とした場合には，各単語の1回の生起が1つのイベントである．たとえば，回答テキストが,「道路 の 早期 の 開通 を 願う」という場合には，7回のイベントがある．そして，回答の属性値が x であるとすると，freq$(x,$の$)$ は +2 され，それ以外については，+1 される．一方，回答クラスタを指標とした場合には，この回答が「あるクラスタAに属する」と考えるので，回答を単位とした1回のイベントがある．このときには，freq$(x,$クラスタA$)$ が +1 される．したがって，イベントの単位に応じて，頻度の計数方法は異なる．つまり，単語をイベントの単位にしたときには，単

(注3)：本稿では,「単語」と「形態素」とを区別せずに用いる．回答テキストの形態素への分割には茶筌 (http://chasen.aist-nara.ac.jp/hiki/ChaSen/) を利用した．

語の出現ごとに計数し，回答をイベントの単位にしたときには，各回答ごとに頻度を計数する．

イベントの単位をどうするかは，アンケートを分析するときに重要である．たとえば，単語を利用したときには，前節の例のように，各属性値に特徴的な単語リストから，大まかな印象をつかむことができるし，また，既存ツールを利用して単語(形態素)を得られるという利点がある．しかし，形態素を利用した場合には，たとえば，「必要」だけが単独で抽出された場合，それが「必要/だ」なのか「必要/ない」なのかが判断できない．これは，形態素という単位と分析に必要な単位とが一致しない場合があることを示している．このような不一致がないような単位を研究することは重要である．その候補としては，単語 n-gram [8] などが考えられるが，それを追及するのは今後の課題である．

一方，もっと大きい (最大の) 単位として，回答自体を使うと，今度は，各回答は具体的であるため，その回答の意味を把握することはできるが，その各回答からアンケート全体の概要を把握することが難しいという問題がある．

それに加えて，頻度を計数するときの問題がある．つまり，回答テキストは，互いに，字面が完全一致するということは，あまりないので，字面が完全一致することをもって，回答テキストが同値である (同一クラスタに属する) とすると，各回答は互いに異なることがほとんどである．そのため，極端な場合には，各回答クラスタ y は，ただ 1 つの回答からなることになる．すなわち，全ての y について，$f(x,y)=1$ である．これは，全回答クラスタが等確率であることを意味するので，これでは，確率的な手段により，特徴的な回答クラスタを抽出することはできない．

そのため，字面の完全一致以外の方法により，回答の同値性を判定し，適切な回答クラスタを作成する必要がある．ここでの同値性は，回答の意味的な同値性を目標とするため，それを達成するのは現時点では困難であるが，次節において，本稿で利用したクラスタリング法について述べる．

以上のように，回答指標に回答クラスタを利用した場合において，イベントの同値性の判定に問題があることは，明らかであるが，回答指標として，単語を利用した場合にも，同値性の判定に問題がある．本稿においては，単語の字面が一致すれば，それらは同値であるとして，頻度を計数した．たとえば，上例では，$f(x,\text{の})$ を+2 する，ということである．しかし，単語においては，多義性や同義性の問題がある [9]．これらへの対処も今後の課題である．

4. 回答のクラスタリング

4.1 クラスタリングの方法

回答のクラスタリングは，回答テキスト t 中の単語 w の出現頻度 $\text{freq}(t,w)$ に基づき (5) 式で推定した確率 $p(t,w)$ を利用する．つまり各回答 t を，単語の確率ベクトル $[p(t,w_1),p(t,w_2),\ldots]$ で表現する．さて，回答テキスト $T=\{t_1,t_2,\ldots\}$ が，k 個の排反なクラスタ $\{T_i|1\leqq i\leqq k\}$，ただし $T=\cup_{i=1}^{k}T_i \wedge T_i\cap T_j=\phi$ に分割されるとき，相互情報量は以下のように変形できる．

$$I(T;W) = I_k(T;W)+D_k(T;W) \tag{6}$$

$$I_k(T;W) = \sum_{i=1}^{k} p(T_i)I(T_i;W) \tag{7}$$

$$D_k(T;W) = \sum_{i=1}^{k} p(T_i)D(p_i(w)||p(w)) \tag{8}$$

$$I(T_i;W) = \sum_{t\in T_i,w} p_i(t,w)\log\frac{p_i(t,w)}{p_i(t)p_i(w)} \tag{9}$$

ただし，$p(T_i)=\sum_{t\in T_i}p(t)$, $p_i(t,w)=\frac{p(t,w)}{p(T_i)}$ である．

$I_k(T;W)$ は，各回答クラスタ T_i 内に局所的に定義される確率 $p_i(t,w)$ により計算される相互情報量 $I(T_i,W)$ の平均値である．$I(T_i,W)$ は，T_i 中の各回答の単語分布が似ているときに小さくなるので，それが小さいほど，まとまりの良いクラスタと言える．一方，$D_k(T;W)$ は，各クラスタ T_i での単語分布 $p_i(w)$ と全体の単語分布 $p(w)$ のダイバージェンス $D(p_i(w)||p(w))$ の平均値である．$D(p_i(w)||p(w))$ が大きいほど，特色のあるクラスタと言える．$I_k(\cdot)$ と $D_k(\cdot)$ は，郡内分散と郡間分散 [10] に対応すると言える．したがって，k 個のクラスタに分けるときには，$\min I_k(T;W)$ (それと等価だが $\max D_k(T;W)$) になるようにクラスタリングすれば良い．なお，$\max D_k(T;W)$ と等価な目的関数が，文献 [11] ～ [14] で提案されているが，本稿での導出方法は，これらとは異なるものである．

本稿の実験において生成されたクラスタを観察したところ，話題としてはまとまったクラスタが得られているようではあるが，まだ，意味的な同値性としては，改善の余地が大きいことが分かった．その改善方法としては，クラスタリングの素性を良いものとすることが有効であると考える．現在のところ単語しか利用していないので，これをもっと良いものにする必要がある．また，クラスタ数の決定という問題もある．文献 [14] では AIC [15] を利用しているが，本稿の実験では，我々の観察では，AIC ではクラスタの粒度が大きいように思えたため，結局は，クラスタ数を人手で指定した．

以上のような問題があるが，回答のクラスタリングは，冗長な回答群を 1 つにまとめるためには，必要なことなので，引き続き検討していきたい．

4.2 クラスタ内の回答や単語の重要性

回答クラスタ内の回答や単語の重要性も相互情報量に基づいて求める．(6) 式から，相互情報量 $I(T;W)$ に対するクラスタ i における寄与分は，$J(T_i;W)=\sum_{i=1}^{k}p(T_i)\{I(T_i;W)+D(p_i(w)||p(w))\}$ である．これより，回答については，

$$J(T_i;W)=\sum_{t\in T_i}\nu(t;W) \tag{10}$$

であるので，$\nu(t;W)$ を回答 t の重要度とする．一方，単語については，$J(T_i;W)=\sum_w\sum_{t\in T_i}p(t)d(w|t)$ より，

$$\omega(w|T_i)=\sum_{t\in T_i}p(t)d(w|t) \tag{11}$$

を重要度とする．これは各回答における単語 w の特徴度の重み付き和である．

表 2　回答テキストが得られた地域とメディアと回答件数
Table 2　Frequency table of response texts

メディア	オープンハウス	アンケート	その他	計
地域 (τ の平均)	(.400)	(.357)	(.243)	
都筑区 (.238)	383	76	31	490
他横浜市 (.205)	291	938	99	1328
青葉区 (.150)	169	109	32	310
緑区 (.144)	90	57	29	176
他都道府県 (.136)	77	0	10	87
他神奈川県 (.128)	67	0	22	89
不明	161	0	57	218
計	1238	1180	280	2698

表 3　回答指標ごとの相互情報量
Table 3　Mutual information

	単語	回答クラスタ
メディア	0.214	0.456
地域	0.208	0.285

5. 実　　験

5.1 デ　ー　タ

提案分析手法の適用例としては,「(仮称) 横浜環状北西線」についての自由回答アンケートの回答テキスト(注4)を利用した．これは，道路建設に関してパブリック・インボルブメント (PI) の一部として行なわれたものである．この回答テキストにおける回答者属性は,「メディア」と「地域」である．それらと回答件数との関係を表 2 に示す．ここで,「メディア」とは，回答者の意見が収集されたメディアであり，それらには「オープンハウス（市民が立寄り，意見できるコーナー）」「アンケート (横浜市内での層化無作為抽出)」「その他（はがき，電話，Web 等）」がある．また,「地域」としては北西線のルートに特に関係する 3 区である「都筑区」「青葉区」「緑区」および「その他の横浜市」「その他の神奈川県」「その他の都道府県」がある．

5.2 分 析 概 要

5.2.1 属性の比較

3. で述べたように，回答指標としては,「単語」と「回答クラスタ」を利用する．このそれぞれを利用した場合の相互情報量を表 3 に示す．なお,「地域」については，地域が「不明」な回答を除外して各種の分析をした．

表 3 より，どちらの指標においても「メディア」の方が「地域」よりも相互情報量が高い．このことから，メディアの方が地域よりも回答指標に反映されやすいと推測できる．しかし，メディアによる意見の違いよりも，地域による意見の違いの方が，特に，道路建設のような地域に直結する事柄においては，大きいと思われるので，メディアの方が反映するというのは不思議である．そのため，その要因を考察した結果，以下のような分析上の問題点が明らかになった．

a)　回答者の受け取る情報の違い

分析対象である PI の関係者からのコメントによると，オープンハウスとアンケートとその他のメディアでは，回答者が回答テキストを書くにあたって参考とする情報 (すなわち図 1 の Z) が異なる．つまり，オープンハウスでは，そこに立寄った市民が，そこでの提示物に影響されて回答テキストを書くのであるが，アンケートでは，アンケート質問用紙に答えるとともに，自由回答アンケートも書く．また，Web 等では，PI のホームページを参考にして回答テキストを書く．したがって，これらが相互情報量に影響している可能性がある．

通常の社会調査では，全回答者が同一の質問事項に回答するので，Z は全回答者に対して一定であるとみなして良い．しかし，PI では，様々な方法を利用して様々な立場の人から様々な意見を得ることを目的の一部としているため，Z が回答者により変化することは必然である．したがって，Z と X との交互作用を考える必要がある．これは今後の課題である．

b)　テキストの電子化の影響

回答テキストを電子化する際に，メディア毎に電子化をした人が違う場合には，電子化のための書き起しに個性がでる可能性が高い．これにより，意見は同じであっても，回答テキストの書き起しの個性により，字面が違うために，メディアが異なると異なる分布をする可能性があり，実際に，そのような傾向があるようであった．この影響の除去も今後の課題である．

以上のような分析上の問題が考えられるので，メディアの違いが回答指標に影響すると結論付ける前に，より詳しく調べる必要がある．

5.2.2 属性値の比較

表 2 では,「メディア」と「地域」の属性値を，2 つの指標における寄与率 τ の平均によりソートしてある．これによると，メディアについては,「オープンハウス」と「アンケート」はほぼ同等，また，地域については,「都筑区」と「その他の横浜市」が他より高く，その他は，ほぼ同等の寄与率である．ここで，メディアについて両者が同等であるのは，回答件数がほぼ同じことから説明できる．次に，地域については，この 2 つの地域は，それぞれ，回答の度数も大きい．また,「都筑区」は「その他の横浜市」より度数は小さいが,「都筑区」が道路建設に関わりの深い地域であることから，特色のある回答が多く，そのことが，寄与率が大きいことに反映していると考える．

5.2.3 指標値の比較

各指標ごとに指標値を特徴度 d でソートして比較する．我々は，分析上，メディアの違いよりも地域の違いに興味があるので，以下では，主に，地域による違いについて観察する．詳細な評価は今後の課題である．

a)　単　　語

上位の内容語のみを調べると，ルートに特に関係する 3 区については，都筑区については,「都筑，便利，ルート，インターチェンジ，なる，よい，心配」があり，青葉区は「青葉，インターチェンジ，必要，行く，羽田，横浜，鶴見川」があり，緑区は「緑，地下鉄，青砥，地下，団地」などがある．一方，その他の横浜市については,「渋滞，アンケート，完成，検討，欲しい，考え」などがあり，その他の神奈川県は「保土ヶ谷，川崎，バイパス，良い，いる，混ん」，その他の都道府県は,「保

(注4)：http://www.yokohama-nwline.jp/ref_pi/

土ヶ谷，バイパス，高速，北西，東名，東京，混ん」などがある．これらより，ルートからの地理的な距離と上位の単語とに強い関係があると推測できる．

b) 回答クラスタ

回答クラスタについても同様に，地域による違いがわかる．興味深い例としては，道路建設に対する反対意見で，ルートにあたりそうな地域の人は，自分の地域がルートにあたることが主要な反対原因であるのに対して，ルートから外れている人については，財政事情の観点等から反対していることが分かることである．

5.2.4 回答のクラスタリングについて

回答のクラスタリングについては，4.1で述べたように，まだ，意味的な同値性としては，改善の余地が大きいことが分かった．4.2で述べた，クラスタ内での回答や単語の重要性については，$\nu(\cdot)$は，長めの回答を優先する．これは，(3)式における$p(\cdot)$の影響である．また，上位の単語については，適切に回答クラスタの内容を表現しているようである．たとえば，回答件数44のクラスタにおいて,「渋滞, の, 保土ヶ谷, 緩和, バイパス, する, 解消, 完成, !, 思う, も, 期待, で, より, 混雑, 北西, ＢＰ, 線, 向上, は, 高, 性, が, と, さ, 横浜, . , ある, 東名, 利用」が上位30位であった．これより,「道路完成による渋滞緩和を期待している」回答からなるクラスタであることが予想でき，それは，実際に，回答を読むことにより，確認できる．しかし道路建設に賛成な回答だけでなく反対も混じってはいる．

6. 関連研究

本稿での主要な貢献は，自由回答アンケートの情報構造を示す(図1)とともに，そこから，情報理論，特に，相互情報量を利用することにより，自由回答アンケートの分析手法が導出可能なことを示したことである．なお，本稿で述べた個々の分析法には，それぞれの目的に対応する別の手法がある場合がある．たとえば，文献[8]では，回答のクラスタリングや特徴的な回答や単語の抽出についてが述べられている．また，対応する手法が既存文献に明示されていない場合でも，本稿で述べる分析と同等な分析をする別の方法はあるだろう．更に，本稿で述べる分析方法自体についても，それと等価な分析法が既出な場合には，それを示した．しかしながら，これら既存の分析方法は，その場の目的に応じて，その都度作られ適用されたものである．それに対して，本稿では，一連の分析方法が，情報理論的な観点から統一的に導出できることを示した．

また，本稿で提案した図1の自由回答アンケートの確率的な構造では，回答指標Yが明示的に分離されている．このことは，回答テキストを分析するためには，その観点となるYが重要であるという我々の立場を示している．すなわち，回答テキストを分析するには，分析の目的に応じた観点を利用する必要があり，我々は，これまでにも，回答に含まれる要求の認定が，そのような観点として重要な1つであることを示してきた[1]．回答指標Yは，そのような観点を分析に明示的に取り込むための変数である．一方，文献[8]や[16]では，本稿での分析と同様に，単語や回答テキスト本体について，各属性値に特徴的なものを抽出する方法等は示されているが，その他の分析の観点を回答の分析に利用する方法については示されていないし，また，単語や回答テキスト本体による回答の分析が，実際には，回答指標からの回答の分析という概念に統一可能なことも把握されていない．したがって，本稿において，回答指標を，他の要素から分離し，1つの変数として抽出したことは，今後，自由回答アンケートを分析する際に，様々な観点を導入できる枠組を提案したという意義がある．我々は，今後，これまでと同様に，そのような観点を，実際に，自由回答アンケート分析に適用するために必要な言語分析などを研究する予定である．

7. おわりに

本稿では，自由回答アンケートに関わる確率的な構造を示すとともに，そこから，情報理論，とくに，相互情報量を利用することにより，自由回答アンケートの分析手法が統一的に導出可能なことを示した．また，その分析手法を利用することにより，フェイスシートの回答者属性が，分析の観点である回答指標に反映される程度を調べることができることを述べた．更に，提案手法を適用する上で，残されている課題を述べた．

謝辞

パブリックインボルブメントに自然言語処理技術を活用する勉強会における議論が参考になった．ミーティングに参加した，乾孝司，奥村学，落谷亮，高橋和子，高村大也，庭田美穂の各氏に感謝する．

文　献

[1] 大塚, 内山, 井佐原：“自由回答アンケートにおける要求意図判断基準”, 自然言語処理, **11**, 2, pp. 21–66 (2004).
[2] 辻, 有馬：“アンケート調査の方法”, 朝倉書店 (1987).
[3] 酒井：“アンケート調査の進め方”, 日経文庫, 日本経済新聞社 (2001).
[4] 林：“質問紙の作成”,「心理学研究法 9 質問紙調査」村上・続編, 4章, P107-137, 東京大学出版社 (1975).
[5] T. M. Cover and J. A. Thomas: “Elements of Information Theory”, John Wiley & Sons, Inc. (1991).
[6] T. Dunning: “Accurate method for the statistics of surprise and coincidence”, Computational Linguistics, **19**, 1, pp. 61–74 (1993).
[7] P. Resnik: “Selection and Information: A Class-Based Approach to Lexical Relationships”, PhD thesis, University of Pennsylvania (1993).
[8] L. Lebart, A. Salem and L. Berry: “Exploring Textual Data”, Kluwer Academic Publishers (1998).
[9] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas and R. A. Harshman: “Indexing by latent semantic analysis”, Journal of the Society for Information Science, **41**, 6, pp. 391–407 (1990).
[10] 田中, 脇本：“多変量統計解析法”, 現代数学社 (1983).
[11] L. D. Baker and A. K. McCallum: “Distributional clustering of words for text classification”, SIGIR'98 (1998).
[12] H. Li and N. Abe: “Word clustering and disambiguation based on co-occurrence”, COLING-ACL'98 (1998).
[13] I. S. Dhillon, S. Mallela and D. S. Modha: “Information-theoretic co-clustering”, KDD'03 (2003).
[14] 高村, 松本：“文書分類のための共クラスタリング”, 情報処理学会論文誌, **44**, 2, pp. 443–450 (2003).
[15] 坂元, 石黒, 北川：“情報量統計学”, 共立出版株式会社 (1983).
[16] 大隅：“調査における自由回答データの解析”, 統計数理, **48**, 2, pp. 339–376 (2000).