論文

データマイニング技法を用いた診断アンケート支援システム
Diagnostic Questionnaire Supporting System Using a Data-Mining Technique

田中 猛彦 *　田中 康幸 †　中川 優 *　小倉 光博 ‡　板倉 徹 ‡

Takehiko TANAKA　Yasuyuki TANAKA　Masaru NAKAGAWA
Mitsuhiro OGURA　and　Toru ITAKURA

質問に答えると診断結果が得られる「診断アンケート」は，実施者・回答者それぞれにメリットがあり，インターネットで広く実施されている．結果スコアを計算するには，あらかじめ重み付けスコアを設定する必要があるが，精密で根拠のあるスコア設定は難しい．そこで本論文では，回答データベースに基づいて重み付けスコアを自動的に修正し，常に最適な設定で結果スコアを計算する手法を提案する．自動修正には，データマイニング技法の一つである相関関係分析を使用している．本手法を脳出血の危険度判定アンケートに適用し，有効性を検証した．13 件の脳出血発症者を含む 713 件の回答に対して，年齢と結果スコアに関する散布図を作成すると，加齢により結果スコアが上昇するような散布図が得られた．また職業などの質問にも，適用により重み付けスコアが割り当てられ，本手法を用いることで知識発見の効果も期待できることがわかった．

Diagnostic questionnaires come into wide use in the Internet. The conductor of a questionnaire has to arrange the weighted score for each option, before deriving the resulting scores from the replies. It is, however, difficult to assign accurate, well-founded weighted scores to all the options. In this paper, we propose a method for calculating the weighted scores according to the reply database of the questionnaire, for the purpose of presenting the optimum resulting score at any time. We adopt a correlation analysis, one of the data-mining techniques, to modify the weighted scores. For verifying the validity, we apply the method to a questionnaire on the cerebral hemorrhage. There were replied 713 answers including 13 ones of those who experienced the cerebral hemorrhage. As a result of using the proposed method, the scatter graph of the age and the resulting score is drawn where the resulting scores increase with aging. Furthermore, several questions such as the one about the occupation are judged to be correlated and the options of these questions are weighted, which means the detection of the novel knowledge.

キーワード：診断アンケート，データマイニング，相関関係分析，脳出血，ウェブアプリケーション
Diagnostic questionnaire, Data mining, Correlation analysis, Cerebral hemorrhage, Web application

* 和歌山大学システム工学部
Wakayama University, Faculty of System Engineering
takehiko@sys.wakayama-u.ac.jp
† 株式会社日立システムアンドサービス
Hitachi Systems & Services, Ltd.
‡ 和歌山県立医科大学脳神経外科
Wakayama Medical University, Department of Neurological Surgery

1　はじめに

1.1　研究の背景

アンケート調査は様々な場面で行われている．最近は，S-PLUS[1]など良質なアンケート処理ソフトウェアが登場してきたことにより，

[1] 日本語版は数理システムが開発している．http://www.msi.co.jp/splus/ が詳しい．

大規模なアンケートの実施・分析を簡単に行えるようになってきた．またインターネットの普及や通信・放送の融合によるネットワークの高度化により，双方向での情報発信が容易になり，従来は郵送などで実施されていたアンケートを，双方向性ネットワークの代表例といえるインターネットを利用して行うケースが増えている[1][2]．インターネットを使った調査には，調査期間の短縮，低費用での実施，場所の制約を解消できるなどの利点があり，今後も，紙媒体によるアンケートから，インターネットを使ったアンケートへと移る傾向が続くと思われる．実際，インターネット上でのアンケートを支援するために，アンケートの作成，回収，集計を支援するサービスが各所で行われている．しかしアンケートには，世論調査のような一般の人を対象に広く調査を行うアンケート，顧客満足度調査のような，企業がユーザに対してデータ収集のために行うアンケートなど様々な種類があり，適切な支援を行うには，実施するアンケートに合わせた支援システムが必要となる．本研究では，回答者が質問に答えることにより診断結果が得られるアンケートを「診断アンケート」と呼ぶことにし，注目することにした．

1.2　診断アンケート

この節では，本研究で対象とする診断アンケートの使用方法，意義，およびアンケート実施の留意点について述べる．

診断アンケート実行の流れを簡単に述べる．まず実施者はアンケート文を作成する．このとき，質問文や選択肢だけでなく，選択肢ごとに点数 (本論文では「重み付けスコア」と呼ぶ) を設定しておく．そして回収した回答に対して，設定に基づいた結果スコアを算出し，この結果スコアに基づき診断結果を回答者に返す[2]．結果スコアの算出方法として，本研究では極めて単純であり，かつ広く利用されている「選んだ各選択肢に割り当てられている重み付けスコアの和を結果スコアとする」方式を採った．

通常のアンケートでは，実施者は回答データを得ることができるが，回答者は統計情報を得るくらいのものであったため，回答を多く集めるためには抽選による懸賞をつけるなどしなければならなかった[3]．診断アンケートでは，回答者も診断結果を得るというメリットがあるために，回答を集めやすいという特徴がある．診断アンケートは雑誌やインターネット上で広く用いられているアンケート形式の一つであり，具体的には危険度判定のような健康診断や，性格診断，理解度判定など，回答者の現在の状況を判断するのに使われている．特に医療関係での効果が期待できるアンケートである．

診断アンケートを実施するにあたり，次の 2 点に留意する必要がある．

1. **重み付けスコアの決定**: 既に述べたように，重み付けスコアは診断アンケートにおいて非常に重要な作業である．この設定が間違っていた場合，正しい診断結果を回答者に返すことができなくなってしまう．しかしアンケート作成時であれ，回収・電子化の後であれ，正しい重み付けスコアを設定するというのは困難な作業である．

2. **リアルタイム性**: 診断アンケートは，アンケート回答時の現状を診断するものであり，そのため回答から診断結果を得るまでの期間が短いほど，回答者にとって有益となる．紙媒体での調査でよく見られるような，回答を実施者に送付し，1 か月後に診断結果が得られるようなシステムでは，回答者が有益な診断結果を得られない可能性がある．

これらを満足し，かつアンケートの実施者も回答者も利用しやすい方式として，Web アプリケーションによる実装が考えられる (図 1)．アンケート実施者はブラウザを使って，Web サーバ上に置かれたアンケート作成支援システムにアクセスし，質問文の作成や重み付けスコアの設定を行う．作成された質問

[2] 紙媒体によるアンケートでは，質問文の印刷，回答の電子化，診断結果の打ち出しなどの作業も必要となる．

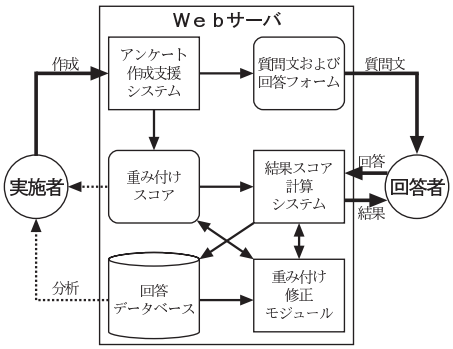

図 1　Web アプリケーションによるアンケート支援システム

文から，Web サーバ上に HTML (HyperText Markup Language) 文書として回答フォームが自動生成される．回答者もまたブラウザを通じて Web サーバにアクセスし，生成された回答フォームを見てアンケートに答える．回答者がフォーム提出ボタンを押すと，Web サーバは回答データをデータベースに格納し，重み付けスコアと回答内容に応じて結果スコアを計算して，診断結果を回答者に提示する．

データベースに蓄積されていく回答は，後に統計処理を通じて，回答者集団の傾向や関心などを知ることが可能となる．その処理方法として近年，データマイニング技法が注目を集めている[4][5][6]．データマイニングの代表的な手法として，相関関係分析，時系列パターン分析，クラスタリングなどが知られているが，本研究では，回答データベースに対して相関関係分析の手法を用いて質問と結果スコアの関係を求め，重み付けスコアの修正を試みている．

1.3　研究の目的

本研究では Web 診断アンケートにおいて，実施者を支援するシステムの開発を行うと同時に，前節で述べた「重み付けスコアの決定」と「リアルタイム性」がともに有効に機能するような方式を検討している．

診断アンケートではアンケート作成時の重み付けスコア設定に大きな手間を要するが，実施者が初期設定の段階から，正確な重み付けをすることは困難である．そこで本研究では，過去に回収した回答データベースから，相関関係分析により重み付けスコアを自動的に変更することにした．相関関係分析を用いる場合，何と何との相関を求めるかが重要である．提案する手法では，質問を，明らかに診断内容と関係があり重み付けスコアを初期状態から変更することのない「スコア固定の質問」と，診断内容と関係があるかが不明 (あるいは，関係を調べてみたい質問) であり重み付けスコアの修正対象である「スコア可変の質問」とに分け，スコア固定の質問のみの結果スコアと，スコア可変の各質問との相関を求める．この定式化は 2 章で述べる．この自動修正機能により，スコア固定の質問に対する重み付けスコアを適切に設定しておけば，常にアンケートを回収した時点での最適な重み付けで，結果スコアを出力できる．

この重み付けスコア修正手法の有効性を確認するため，和歌山県下で実施した脳出血危険度判定アンケートへの適用を試みた．3 章で，その実施内容，本手法適用の概要および結果を示す．

Web アプリケーションにおける「リアルタイム性」とは，回答者が回答を送信してから，どのくらいの時間 (ターンアラウンド時間) で結果を得ることができるかに関わる．ターンアラウンド時間が数秒でも利用者はストレスを感じると言われている．4 章で，重み付けスコアの修正処理の効率化を検討するとともに，結果の意味付けや，悪意ある回答への対処といった，実用性に関しても考察する．

本システムは，Common Gateway Interface (CGI) が利用可能な Web サーバと，著者らが開発しているアンケート作成支援システム ([7][8]でその詳細が述べられており，本論文では立ち入らない) およびスコア計算システムがあれば利用できる．そのため，計算機に関する高度な技術を持たない人でも容易にアンケートを設置でき，このことにより，手軽にアン

ケートを実施できるようになることが期待される．

2　診断アンケートの定式化

「選択肢ごとに設定された重み付けスコアの総和により回答者の結果スコアを計算する」方式のアンケートについて，2.1 節でそのスコア決定方法を数学的に記述する．2.2 節は本論文で提案する数学的枠組であり，相関関係分析に基づき，重み付けスコアを変更する方法を述べている．

2.1　スコア決定方法 — 重み付けスコアが固定の場合

質問，回答の候補: 質問数を n とし，質問 $q_1, q_2, \ldots, q_n$ のそれぞれに対して，回答の候補からなる集合 $D_1, D_2, \ldots, D_n$ が定められているものとする．それらの集合は一般に無限集合でもよい．例えば，自由記述の質問には，任意長の文字列全体からなる集合が対応する．

回答, 回答データベース: 回答者集合を $A = \{a_1, a_2, \ldots, a_m\}$ とする．回答者 $a_i \in A$ のアンケート回答は，n 字組 $(a_{i1}, a_{i2}, \ldots, a_{in})$ で表わされる．ここで各 $1 \leq j \leq n$ に対して，a_{ij} は質問 q_j の回答を表わし，$a_{ij} \in D_j$ を満たすものとする．回答データベースは, m 行 n 列の行列 M により表わされる．ここで，行列 M の (i, j) 成分は a_{ij} である．

重み付け: $\boldsymbol{R}$を実数全体からなる集合とし，各質問 q_j に対して，写像 $f_j: D_j \to \boldsymbol{R}$ がちょうど一つ定まるものとする．すなわち，質問 q_j の各回答 $d_j \in D_j$ に対して，$f_j(d_j)$ は実数を与える．この値を，回答 d_j の「重み付けスコア」と呼び，f_j を，質問 q_j の「重み付け関数」と呼ぶ．

結果スコア: 以上の準備のもとで，回答者 a_i の結果スコアは

$$R(a_i) = \sum_{j=1}^{n} f_j(a_{ij}) \tag{1}$$

と表現される．

2.2　スコア決定方法 — 重み付けスコアが変動する場合

2.2.1　スコア決定の概要

質問: 質問数を n とし，集合 $\{1, 2, \ldots, n\}$ の直和分割となるような 2 つの集合 B, V を定める．すなわち，$B \cap V = \emptyset$，$B \cup V = \{1, 2, \ldots, n\}$ が成り立つものとする．質問 q_j を，$j \in B$ のとき「スコア固定の質問」，$j \in V$ のとき「スコア可変の質問」と呼ぶ．

回答の候補，回答，回答データベース: 2.1 節と同じである．

重み付け: 重み付けに変動がない場合と同じく，各質問 q_j に対して，写像 $f_j: D_j \to \boldsymbol{R}$ をちょうど一つ定める．$j \in V$ に対する f_j は，回答データベースの内容により変化するが，スコア決定時にはちょうど一つ定まるものとする．

結果スコア: 回答者 a_i の回答に対する「基本スコア」を

$$R_b(a_i) = \sum_{j \in B} f_j(a_{ij}) \tag{2}$$

により計算する．スコア可変の質問ごとに，基本スコアとの相関比を求め (2.2.2 節)，その値により重み付けスコアを変更して (2.2.3 節) から，回答者 a_i の回答に対する「変動スコア」を

$$R_v(a_i) = \sum_{j \in V} f_j(a_{ij}) \tag{3}$$

により計算する．最終的に，回答者 a_i の結果スコアは

$$R(a_i) = R_b(a_i) + R_v(a_i) \tag{4}$$

$$= \sum_{j=1}^{n} f_j(a_{ij}) \tag{5}$$

で求められる．

スコア決定の大まかな流れを図 2 に示す．

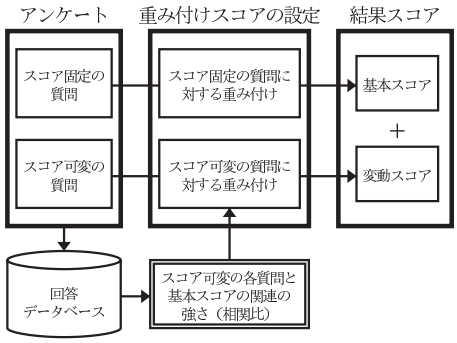

図 2　スコア決定の流れ

相関比はスコア可変の質問ごとに求めるが，重み付け変更においては相関比の合計に対する割合に応じて，重み付けスコアを割り振っている．

2.2.2　相関比の計算

スコア可変の質問 q_j $(j \in V)$ を任意に固定する．この質問と基本スコアとの相関を調べるにあたり，単純な相関係数を求めることは，D_j が一般に定量的な集合ではないため，意味を持たない．ここでは，全回答者に対して，回答した選択肢 (D_j の要素) ごとにグループ分けを行い，このグループ化に基づき級内変動，級間変動，および相関比を計算することにした．これらの一般的な計算方法は[9]が詳しい．

以下では簡単のため，次の 2 つの仮定をおく．一つは，無回答を認め，その場合には常にその重み付けスコアを 0 とする．言い換えると，回答の候補の各集合 $D_1, D_2, \ldots, D_n$ は，無回答に対応する要素 λ を有し，各 $1 \leq j \leq n$ に対して常に $f_j(\lambda) = 0$ が成り立つものとする．もう一つの仮定として，スコア固定の各質問の重み付けスコアはいずれも非負とする．これは，スコア可変の質問に対する各回答のスコア割り振りに，回答ごとの平均基本スコアを用いるためである．スコア固定の質問の重み付けスコアで負のものがあった場合，その質問のすべての回答の候補にいわばかさ上げすることで，いずれも非負にすることが常に可能である．

準備: 回答データベースの中で，質問 q_j に対して $d_j \in D_j$ (ただし $d_j \neq \lambda$) を回答した回答者の集合を $A(q_j{:}\, d_j)$ と表記する．文脈から質問 q_j が一意に定まる場合は $A(d_j)$ と略記する．さらに，D_j の各要素の中で，それを回答した者が存在するものからなる集合を $\tilde{D}_j$ と書く．すなわち $\tilde{D}_j = \{d_j \mid d_j \neq \lambda, A(d_j) \neq \emptyset\}$ である．D_j が無限集合であっても，回答者数が有限であれば，$\tilde{D}_j$ は有限集合となる．この事実と，重み付けスコアの修正は $\tilde{D}_j$ の各要素に対してのみ行えばよいことを合わせると，処理の停止性が保証される．

回答者からなる空でない集合 $A' \subset A$ について，A' に属する回答者の基本スコアの平均を $\overline{R_b(A')}$ と表記し集合 A' の「平均基本スコア」と呼ぶ．これは

$$\overline{R_b(A')} = \sum_{a_i \in A'} R_b(a_i) \,/\, |A'| \qquad (6)$$

により求められる．ここで有限集合 X に対して $|X|$ はその要素数を表す．以下で用いるのは，全回答者の平均基本スコア $\overline{R_b(A)}$ と，スコア可変の質問 q_j に対して d_j を回答した回答者の平均基本スコア $\overline{R_b(A(q_j{:}\, d_j))}$ (もしくは $\overline{R_b(A(d_j))}$．これを，「回答 d_j の平均基本スコア」と呼ぶ) である．

級内変動: 質問 q_j の級内変動 $S_{wc}(q_j)$ とは，各回答者 a_i の基本スコアが，回答 a_{ij} の平均基本スコアからどのくらい変動しているかを表わす値である．すなわち，

$$S_{wc}(q_j) = \sum_{i=1}^{m} \left(R_b(a_i) - \overline{R_b(A(a_{ij}))} \right)^2 \qquad (7)$$

で計算される．

級間変動:質問 q_j の級間変動 $S_{bc}(q_j)$ とは，各回答者 a_i の回答 a_{ij} の平均基本スコアが，全回答者の平均基本スコアからどのくらい変動しているかを表わす値である．すなわち，

$$S_{bc}(q_j) = \sum_{i=1}^{m} \left(\overline{R_b(A(a_{ij}))} - \overline{R_b(A)} \right)^2 \qquad (8)$$

で計算される．

相関比: 相関比とは，級内変動と級間変動の合計に対する級間変動の割合であり，級内変

動が小さいとき，もしくは級間変動が大きいときに，その値が大きくなる．基本スコアに対する質問 q_j の相関比 η_j^2 は，$S_{bc}(q_j)$，$S_{wc}(q_j)$ の少なくとも一方が 0 でないとき，

$$\eta_j^2 = S_{bc}(q_j)/(S_{bc}(q_j) + S_{wc}(q_j)) \quad (9)$$

で求められる．以下ではこれを，「質問 q_j における相関比」と呼ぶ．全回答者の基本スコアが一致する[3]とき，かつそのときに限り $S_{bc}(q_j) = S_{wc}(q_j) = 0$ となるが，このときは $\eta_j^2 = 0$ と定める．

式 (7) および式 (8) より，$S_{wc}(q_j) \geq 0$，$S_{bc}(q_j) \geq 0$ であり，これらと式 (9) から，$0 \leq \eta_j^2 \leq 1$ であることがわかる．$\eta_j^2 = 1$ になるのは $S_{wc}(q_j) = 0$ と $S_{bc}(q_j) > 0$ が同時に成り立つときかつそのときに限られ，これは，回答の候補 ($\tilde{D}_j$ の要素) ごとに，その回答者の基本スコアが同じ値であり，かつ回答の候補ごとにはその値に何らかの差異があることを意味する．$\eta_j^2 = 0$ になる必要十分条件は $S_{bc}(q_j) = 0$ であり，これは，回答の候補ごとの平均基本スコアがすべて一致することに対応する．

2.2.3　重み付けスコアの修正

前節の方法で，スコア可変の各質問における相関比を計算した後，その相関比を用いて重み付けスコアの修正を行う．具体的には，まず相関比から各質問へのスコアを割り当て，次に質問ごとに，各回答の候補に重み付けスコアを設定する．

ここで 2 つのシステムパラメータ α, β を導入する．α は関連度の低い質問に対して重み付けスコアを割り当てないようにするためのもので，$0 \leq \alpha \leq 1$ である．スコア可変の質問が多いほどこの値を小さくすべきであり，例えば $\alpha = 1/|V|$ (スコア可変の質問数の逆数) が適切と思われる．β は修正後のスコア可変の質問に割り当てる重み付けスコアの総計の最大値で，$\beta \geq 0$ である．この値としては例えば，「スコア固定の質問に割り当てる重み付けスコアの総計」を基準にして，その定数倍とするのが妥当であろう．

以下では，少なくとも一つのスコア可変の質問 q_j $(j \in V)$ に対して $\eta_j^2 > 0$ であるものとする．

重み付けを行う質問の判定: スコア可変の質問 q_j $(j \in V)$ に対して，スコア可変の質問全体における相関比の割合 L_j を「関連度」といい，

$$L_j = \eta_j^2 \Big/ \sum_{k \in V} \eta_k^2 \quad (10)$$

により求める．そして $L_j \geq \alpha$ であれば，質問 q_j に対して重み付けスコアを割り当て，そうでなければ重み付けスコアを割り当てない (すなわち，$f_j \equiv 0$ とする)．

重み付けスコアの設定: ここで若干の記法を導入する．まず $\overline{R_b(q_j : d_j)}$ を $\overline{d_j}$ と略記する．次に，$\{\overline{d_j} \mid d_j \in \tilde{D}_j\}$ の中で最小のものを δ_j と記す．

これらを用いて，$L_j \geq \alpha$ を満たす質問 q_j，およびその各回答の候補 $d_j \in \tilde{D}_j$ に対して，新たに割り当てる重み付けスコア $f_j(d_j)$ は，

$$f_j(d_j) = \beta \times L_j \times l(d_j) \quad (11)$$

により設定する．ここで，

$$l(d_j) = (\overline{d_j} - \delta_j) \Big/ \sum_{d \in \tilde{D}_j} \overline{d} \quad (12)$$

である．「2 段階の配分」を行っており，式 (11) において，L_j が最初の段階，$l(d_j)$ が次の段階の配分割合を表わす．このとき，$f_j(d_j) \geq 0$ が常に成立し，かつ質問 q_j ごとに，$f_j(d) = 0$ となるような回答の候補 $d \in \tilde{D}_j$ が存在して $\overline{d} = \delta_j$ を満たす．式 (12) は，2 段階目の配分に平均基本スコアに基づく比例配分を行うが，平均基本スコアが最小の回答の候補は，その重み付けスコアを 0 になるよう，重み付けを減らしていることを意味する．単純な比例配分ではなくこの式を用いるのは，無回答と各

[3] この場合はスコア可変のすべての質問 q_j $(j \in V)$ に対して $\eta_j^2 = 0$ となり，重み付けを変更することができない．このとき，回答者数 m の値が非常に小さい (例えば $m = 1$ のときは常に成り立つ) のでなければ，スコア固定の質問の設定が適切でない可能性が考えられる．

回答との重み付けスコアの差を小さくするためである．

3　脳出血危険度判定アンケートへの適用

前章で述べた，相関関係分析に基づく重み付けスコア修正手法の有効性を検証するため，脳出血危険度判定アンケートへの適用を試みた．以下にその方法および結果を示す．全ての質問文と選択肢，アンケート実施時の重み付けスコアを含め，詳細は[8]に記されている．

3.1　脳出血危険度判定アンケート

適用する脳出血危険度判定アンケートについて説明しておく．このアンケートの目的は，アンケートから回答者の，脳出血を発症する危険度を判定することにある．アンケートの質問項目は，職業，家族構成，食生活など，生活習慣に関する 62 個の質問から構成されている．質問内容の要約を付録にまとめた．米国の調査事例[10]などをもとに，質問項目の作成および重み付けスコアの初期設定を行い，そのアンケートと回答用紙を，紙面により和歌山県下で配布回収した．病院の協力により，過去に脳卒中を発症し現在は回復した 65 名の方からも回答を得ており，回答時の年齢や生活習慣ではなく，発症時 (発症直前) の経験に基づき回答してもらった．

脳卒中には大きく分けて脳梗塞，くも膜下出血，脳出血の症状があり，実際のアンケートでは，これら 3 つの危険度を別々に求め (別々に重み付けスコアを設定し)，回答者に結果を返送した．本論文での適用実験では，脳出血に注目し，その危険度を求めることにした．回答数は 713 件であり，脳出血の発症者 13 件の回答も含まれている．

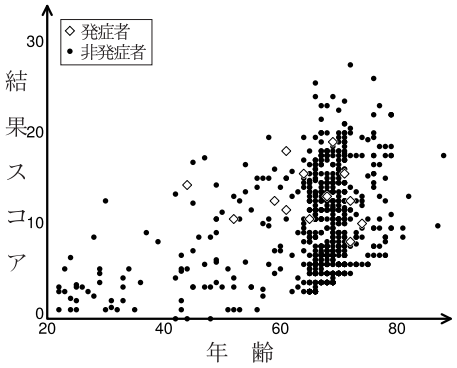

図 3　脳出血危険度判定アンケートの散布図 (アンケート実施時)

3.2　結果スコアの算出 — アンケート実施時の重み付けスコア

まず，アンケート実施時の重み付けスコアを用いて，各回答者の結果スコアを求め，散布図を作成した．その散布図を図 3 に示す．アンケート作成者側としては，発症者と非発症者との間に，結果スコアによる明確な差があり，また年齢が上がるほど脳出血の危険度が高くなることが望ましい．結果スコアによる明確な差があれば，結果スコアを使って回答者を脳出血の危険がある人と，そうではない人に分けることができる．しかし散布図を見ると，結果スコアによる差が多少あるものの，発症者のデータに注目したところでは，発症者でもスコアが低いデータがあり，さらに年齢の増加と共に結果スコアが低くなる傾向も見られた．

3.3　結果スコアの算出 — 提案手法による重み付けスコア

2.2 節で述べた手法を用いて，重み付けスコアの修正を試みた．その際の注意点をいくつか記す．

質問の分類については，以下の通りとした．最初に，数値や文で回答してもらうため単純な統計処理が難しく，重要度の低いと思われ

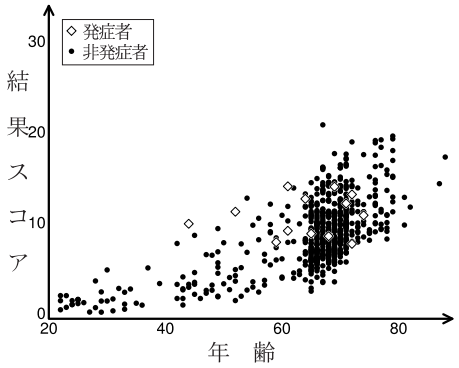

図 4　脳出血危険度判定アンケートの散布図 (提案手法)

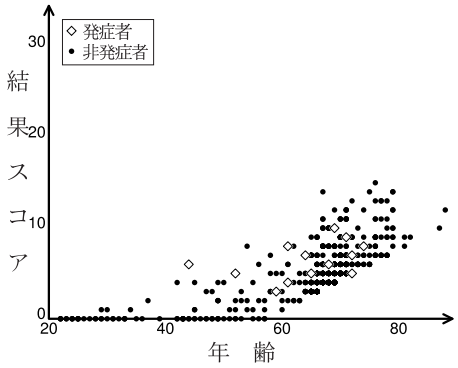

図 5　脳出血危険度判定アンケートの散布図 (基本スコア)

る 5 つの質問を，結果スコアの計算から除外した．次に，脳出血の危険度と明らかに関連があると言われている，年齢，脳卒中の発症経験とその種類，上の血圧 (収縮期血圧) をスコア固定の質問とし，回答項目ごとに前節と同じ重み付けを行った．これらを除く 53 個の質問を，スコア可変の質問とした．

スコア可変の質問の重み付けスコアは，回答ごとに修正するのではなく，まず全回答を対象として重み付けスコアを設定し，それから，その重み付けスコアに基づいて各回答の結果スコアを算出している．重み付けスコア割り当てのためのシステムパラメータについては，$\alpha = 0.02$, $\beta = 61.25$ としている．前述の通りスコア可変の質問数は 53 であるが，その概数 50 の逆数を α とした．また，アンケート実施時の重み付けスコアの合計は 61.25 であり，この値を β とした．

このようにして求めた各回答者の結果スコアについて，その散布図を図 4 に示す．また，比較のため，基本スコア，すなわちスコア固定の質問のみで求めた結果スコア[4]についても，その散布図を図 5 に示す．

3.4　考察

今回の実験において考察すべき点は，「結果スコアの変化」と「関連があると思われる質問の発見」に大別される．以下，それぞれの詳細を述べる．

結果スコアの変化 (1): ここでは各散布図を比較し，全体的な結果スコアの違いについて述べる．図 3 および図 4 を比べてみると，本手法の適用により，発症者のみでも，全体としても，ばらつきが少なくなっているのがわかる．また，発症者であるにも関わらずその年齢の中で結果スコアが最も小さいという例 (図 4 の 72 歳の発症者) もあり，本手法は，発症者と非発症者を明確に区別できるとは言い難い．なお，発症者のみに着目すると，本手法の適用により，52 歳 (左から 2 番目) の結果スコアが相対的に高くなっており，また発症者が上 8 人と下 5 人にグループ分けできるような分布になっている．これについては，「関連があると思われる質問の発見」について述べてから，詳しく考察する．

既に述べたように，加齢に伴い脳出血の危険度が高くなるのが望ましい．本実験においてそのことを定量的に検討するために，回帰直線 (最小 2 乗法による単回帰線) の式を求めた．結果を表 1 に示す．3 種類の散布図，および発症者のみか全員かで，6 つの場合に分けて算出している．X の係数が，加齢による結果スコアの上昇率であることに注意すると，これらの式から，アンケート実施時の重み付

[4] 提案手法で $\beta = 0$ として重み付けを行った場合の結果スコアでもある．

表 1　回帰直線式

発症者・図 3:	$Y = 15.895 - 0.035X$
発症者・図 4:	$Y = 9.134 + 0.034X$
発症者・図 5:	$Y = 0.238 + 0.096X$
全員・図 3:	$Y = -1.754 + 0.182X$
全員・図 4:	$Y = -5.374 + 0.225X$
全員・図 5:	$Y = -6.817 + 0.183X$

(X: 年齢, Y: 結果スコア)

けでは加齢により脳出血発症者の結果スコアが下降する傾向にあること，それに対して本方式ではそれが上昇する傾向にあることが確認できる．次に，全員を対象とした 3 つの回帰式を比較すると，図 4 すなわち提案手法の X の係数が，図 5 すなわち基本スコアのそれよりも大きくなっていることがわかる．これは，提案手法の適用により，基本スコアでは 0 点だった 20〜50 代の回答者に対して，年齢に応じて加点されていることに関係があると思われる．

本論文の提案手法では，スコア可変の質問にあらかじめ重み付けを行っていても，それは重み付け修正に依存しない．一方，実際のアンケートでは，スコア可変の質問のいくつかに対して最初に重み付けが行われていた．この点に注意すると，スコア固定の質問を適切に設定した上で提案手法の適用により，「加齢により結果スコアが上昇する」という，アンケート実施時の予想を反映させた散布図が作成できたということになる．

関連があると思われる質問の発見: 初期設定では重み付けスコアが 0 だったいくつかの質問に対しても，基本スコアとの相関が見られたことから，本手法の適用後には重み付けスコアが割り当てられていた．例えば表 2 のように，職業の質問には，アンケート実施時に重み付けスコアを設定していなかったが，回収データからは基本スコアと相関があると判断し，重み付けスコアが割り振られた．ここで，職業の関連度が高くなった理由として，年齢との相関が考えられる．また，「漁業関連」が最も大きい重み付けスコアとなっているが，

表 2　重み付けスコアの修正例 (職業)

回答	重み付けスコア	平均基本スコア	回答者数
農林関連	0.579	5.380	266
漁業関連	0.990	8.000	1
自営業	0.604	5.536	69
事務職	0.300	3.600	5
会社員	0.016	1.788	33
公務員	0.000	1.688	16
会社役員	0.630	5.700	10
主婦	0.535	5.096	83
無職	0.705	6.180	122
その他	0.332	3.800	55
無回答	0.000	—	53

表 3　重み付けスコアを割り当てた質問

質問 (要約)	関連度
高血圧と言われた	0.142
職業	0.120
飲酒頻度	0.076
下の血圧	0.055
規則的な生活	0.048
かかりつけの医師	0.047
家族構成	0.038
健康番組を信じる	0.034
コレステロールを下げる薬	0.032
家族の人数	0.031
病院に行く	0.031
規則的な食事	0.029
突然の激しいめまい	0.027
ストレス解消法	0.026
ストレスを感じる	0.023
喫煙	0.021

これは回答者数が 1 であり，その基本スコアがたまたま高かったためと推測できる．

すべての質問に対して重み付けスコアが振られたのではなく，例えば，性別に関する質問に対する関連度は 0.002 であり，重み付けスコアを割り当てなかった．関連度が $\alpha = 0.02$ 以上であり，重み付けスコアを割り当てた質問とその関連度を表 3 に示す．

これらのことから，本手法を用いることで，実施者がそれまで考えていなかった事実への手がかりとなり，知識発見へとつながる効果もあることもわかった．

結果スコアの変化 (2)：ここでは特徴的な回答者を取り上げる．図 3〜5 において，最も左にある 2 人の発症者 (44 歳と 52 歳) の結果スコアを比較すると，提案手法では 52 歳発症者の結果スコアが 44 歳発症者より高く，残り 2 つの散布図では逆になっている．全体の中の位置付けとして見ると，52 歳発症者の結果スコアは，提案手法によりいわば押し上げられていることがわかる．

ここで，この 2 人の回答者の回答と重み付けスコアを精査したところ，次の 2 つが判明した．一つは，ともに「高血圧と言われた」の質問に「はい」と回答していることである．この質問はアンケート実施時に重み付けがなされており，「はい」には 6 点が設定されていた．提案手法の適用ではこれが 2.050 点となっている[5]．この重み付けの妥当性を調べるため，「高血圧と言われた」の質問もスコア固定の質問とし，それ以外の条件は変えずに各回答者の結果スコアを求めて，散布図を作成した．その結果，大きく 2 つの領域 (傾きの異なる 2 本の右上がり直線の周辺) に分布し，図 4 の発症者の上 8 人と下 5 人が，そのままその 2 つの領域の上と下に分かれるものになった．実際，「高血圧と言われた」の質問に対して，この上 8 人は「はい」と答え，下 5 人は「いいえ」と回答している．これを採用すると，「高血圧と言われた」の質問が「はい」か「いいえ」かによって，脳出血の危険度が大きく分かれることを意味し，様々な質問への回答をもとに結果スコア (危険度) を求めるものではなくなってしまう．言い換えると，「高血圧と言われた」の質問に対して「はい」の回答に 6 点をつけるのは，提案手法による調査により高すぎたことを示唆している．

もう一つの興味深い点は，アンケート実施時には重み付けがなされなかったが，提案手法では重み付けがなされたいくつかの質問で，44 歳発症者と 52 歳発症者の回答が分かれ，これにより 52 歳発症者の結果スコアに加点されていることである．具体的には，「飲酒頻度」(0.969 点)，「かかりつけの医師」(0.559 点)，「すぐ病院へ行く」(0.253 点) などである．

これらの質問は，ただちに脳出血と関連があるとは断定できず，また回答者の偏り (和歌山の地域性) にもよる可能性がある．実際のところ，アンケートにより脳出血の危険因子を特定するには，より広い地域を対象として実施し，多数の回答者を集める必要がある．むしろ本研究の価値は，図 3〜4，図 4〜5 といった散布図間の比較で気になる点があれば，その回答を調査し，原因となる質問の発見が容易となるツールを開発した点にある．

4　リアルタイムアンケートシステムの実現に向けて

これまで述べてきた手法を用いてリアルタイム診断アンケートシステムを実現するにあたり，いくつか解決すべき問題がある．例えば，「リアルタイム処理のための計算量低減」や「結果スコアの意味付けと提示方法」，「悪意のある回答者への対応」が挙げられる．以下ではこのそれぞれについて検討する．

4.1　リアルタイム処理のための計算量低減

1 章で述べたように，アンケート支援システムを Web アプリケーションとして実装し，回答が登録されるたびに重み付けスコアを更新するようにすれば，常にアンケートを回収した時点での最適な重み付けで，結果スコアが出力できることになる．この点に留意して，回答に対する結果スコアの計算量について考察しておく．

[5] この点数は β に依存している．実際，β の値を約 3 倍にすれば，点数を 6 点にできる．ただしそうすると，重み付けの修正を行った他の回答の候補も同じ比率で上昇することになる．

級内変動を求める 2.2.2 節の式 (7) では，全回答者の基本スコア $R_b(a_1)$, $R_b(a_2)$, ..., $R_b(a_m)$ が必要であり，回答者数 m に比例した時間がかかるように見える．しかしその和を求める順序を変更し，$i = 1, 2, \ldots, m$ の順番ではなく，回答の候補 $d_j \in \tilde{D}_j$ ごとに計算するように変形できる．その上で，一般的な統計処理で分散を効率よく求める方法にならって変形すると，次式が得られる．

$$S_{wc}(q_j) = \sum_{d_j \in \tilde{D}_j} \left(Q(d_j) - \frac{T(d_j)^2}{|A(d_j)|} \right) \quad (13)$$

ただし，$Q(d_j) = \sum_{a_i \in A(d_j)} R_b(a_i)^2$,
$T(d_j) = \sum_{a_i \in A(d_j)} R_b(a_i)$ である．この式により，質問 q_j の回答の候補 $d_j \in \tilde{D}_j$ ごとに，その回答者数 $|A(d_j)|$，基本スコアの総和 $T(d_j)$ および 2 乗和 $Q(d_j)$ を保存しておけば，その質問における級内変動が計算できる．

級間変動に関する式 (8) も同様に

$$S_{bc}(q_j) = \sum_{d_j \in \tilde{D}_j} \frac{T(d_j)^2}{|A(d_j)|} - \frac{T^2}{m} \quad (14)$$

と変形できる．ここで $T = \sum_{i=1}^{m} R_b(a_i)$ である．したがって，上記の値のほかに，全回答者数 m と，全回答者の基本スコアの総和 T を保存しておけば，級間変動が求められる．

$m + 1$ 番目の回答 $(a_{m+1,1}, a_{m+1,2}, \ldots, a_{m+1,n})$ が新たに登録されたとき，$|A|$ および $|A(q_j{:}\, a_{m+1,j})|$ $(1 \le j \le n)$ のそれぞれに 1 を加え，各回答 $a_{m+1,j}$ $(1 \le j \le n)$ に関する基本スコアの総和と 2 乗和，そして全回答者の基本スコアの総和のみを更新すればよい．これにより，回答が一つ増えたとしても，重み付けスコアの修正に要する時間計算量は，回答の候補の数に比例するだけであり，それまでの回答者数に依存しない．この方法では各回答の候補ごとに，回答者数，基本スコアの総和および 2 乗和を格納する領域を必要とするが，実際のアンケートにおいて質問はせいぜい数十問，選択肢は 1 問につき 10 個以内がほとんどであり，回答者が数千〜数万になる可能性があることと比較して，十分小さい領域と言える．

4.2　結果スコアの意味付けと提示方法

リアルタイムアンケートシステムで回答データが追加されるたびに，重み付けスコアを変更するとなると，回答者の結果スコアを単純に比較することはできない．すなわち，$R(a_{100}) = 50$，$R(a_{300}) = 40$ という結果スコアになったとしても，そこからただちに，100 番目の回答者が 300 番目の回答者よりも危険であると判断するわけにはいかない．

3 章で述べた脳出血アンケートでは，結果スコアだけでなく，近い年齢の間で相対的な位置を示すのがわかりやすいことから，図 4 のような散布図上で，回答者の点を強調表示するような方法が有効である．また，回答者全体や発症者の回帰直線も乗せることで，位置がより明確になる．2 次元の散布図による表現方法は，結果スコアと年齢とを結び付けて考えることができるという前提を要するが，脳卒中や生活習慣病など多くの疾病に関する診断アンケートで利用可能と思われる．

4.3　悪意のある回答者への対応

インターネットでリアルタイムアンケートを実施する場合，瞬時に回答が得られるという特徴を利用して，回答の一部またはすべてを変えていきながら何度も試し，良い結果スコアを得ようとする回答者が現れるかもしれない．また，何度も同じ回答を送って，回答データベースの内容を悪くすることもできてしまう．

これらに対して，同一発信者からの短期間の複数リクエストを却下するようにすると，ある程度の悪用を排除できる．ただし，Web サーバから知ることのできる発信者情報は通常，送信元の IP アドレス程度であり，プロキシ

サーバやネットワークアドレス変換 (Network Address Transform, NAT) を経由したアクセスの場合，まれに別の発信者のリクエストも却下してしまう．

これとは別に，全回答数が多くなると，悪意のある回答による結果スコアの影響が小さくなることを用いる方法も考えられる．例えば，あらかじめ数百例のアンケートを実施しておき，それからインターネット上でもアンケートを実施できるようにする．あるいは，回答者は自分がかつて登録した回答内容をいつでも閲覧でき，その時点での回答データベースに基づいて結果スコアや散布図が得られるようにしておく．回答内容を修正して，その修正内容により重み付けを計算し直すような実装も可能である．4.1 節の議論を拡張すれば，回答内容の変更に関しても，重み付けスコアの修正に要する時間計算量をそれまでの回答数に依存しないようにできる．

5　おわりに

本論文では，Web アプリケーションによる診断アンケート支援システムについて，回答データベースに対して相関関係分析を適用することで重み付けスコアを修正する方法を提案し，脳出血危険度判定アンケートに適用してその有効性を検討した．その結果，加齢より危険度が上昇するような散布図が得られたことを確認するとともに，職業などと基本スコアとの相関を見出し，重み付けがなされるといった，新たな発見を得た．

本手法をリアルタイムアンケートとして利用する場合，4.2〜4.3 節で述べた事項以外にも，検討すべき課題が残っている．例えば，システムパラメータ α, β の設定方法や，スコア固定・可変の質問の分類の妥当性などである．今後は，リアルタイムアンケートの運用実験を通じて，それらへの方策を検討していく必要がある．

謝辞

脳卒中アンケートの実施に関しては，プロシード株式会社 前真司企画部長より多くの有益なアドバイスを賜りました．ここに感謝します．

参考文献

[1] 稲石守男; 橋本明宏:「アンケート自動集計システムの構築と運用」, 高エネルギー加速器研究機構 技術部 技術研究会報告集, 5–16, 1999.

[2] 能見正:「双方向性ネットワークを利用した調査手法とその影響」, 郵政研究所月報, No.144, pp.72–97, 2000.

[3] 湯浅秀道:「Internet 調査の方法—Internet 調査の成功のために—」, コンピュータサイエンス, Vol.6, No.1, 1999.

[4] Cheeseman, Peter; Stutz, John: “Bayesian Classification (AutoClass): Theory and Results”, Advances in Knowledge Discovery and Data Mining, AAAI Press, pp.153–180, 1996.

[5] 上田太一郎:「データマイニング事例集」, 共立出版, 192p., 1998.

[6] 上田太一郎:「データマイニング実践集」, 共立出版, 175p., 1999.

[7] 田中康幸; 田中猛彦; 中川優:「データマイニング技法を用いたアンケート支援システムの構築」, 電子情報通信学会技術研究報告, Vol.102, No.603, pp.1–6, 2003.

[8] 田中康幸:「データマイニング技法を用いたアンケート支援システムの構築」, 和歌山大学システム工学研究科修士論文, 2003.

[9] 杉原敏夫; 藤田渉:「多変量解析」, 経済の情報と数理 13, 牧野書店, pp.86–91, 1991.

[10] Wolf, Philip A.; D’Agostino, Ralph B.; Belanger, Albert J.; Kannel, William B.: “Probability of Stroke: A Risk Profile from the Framimgham Study”, Stroke, Vol.22, No.3, pp.312–318, 1991.

付録　脳出血アンケートの質問内容

スコア固定の質問: 年齢; 脳卒中発症経験; 脳卒中の種類 (脳出血, 脳梗塞, くも膜下出血); 上の血圧 (～100，101～120，121～130，131～140，141～150，151～160，161～170，171～180，181～．単位: mmHg).

以下では，アンケート実施時に重み付けスコアを設定していた質問に「∗」をつけている.

提案手法の適用により重み付けスコアを割り当てた質問(関連度で降順): 高血圧と言われた∗; 職業 (農林関連，漁業関連，自営業，事務職，会社員，公務員，会社役員，主婦，無職，その他); 飲酒頻度; 下の血圧∗(～80，81～90，91～100，101～110，111～．単位: mmHg); 規則的な生活; かかりつけの医師; 家族構成; 健康番組を信じる; コレステロールを下げる薬∗; 家族の人数; 病院に行く; 規則的な食事; 突然の激しいめまい; ストレス解消法; ストレスを感じる; 喫煙.

提案手法の適用により重み付けスコアを割り当てなかった質問(関連度で降順): 心臓病疾患経験; 塩辛い物を食べる; 肉類を食べる; 突然の視覚喪失; 突然片側の脱力感・しびれ; 高血圧の薬を正しく服用∗; 高血圧の薬を服用∗; 1 日の喫煙本数∗; 飲酒; 糖尿病と言われた∗; 魚類を食べる; 不整脈; 職場でタバコを吸う人; 仕事や趣味に生きがい; 心筋梗塞∗; 健康番組を見る; 血縁者に高血圧症; 1 回の飲酒量∗(約 1 合; 約 2 合; 3 合以上); 心房細動∗; 夜眠れる; 緑黄色野菜を食べる; 心臓病その他; 健康の話題に敏感; 健康づくりに気をつける; 突然の激しい頭痛; コレステロール・中性脂肪が高い∗; 血縁者にくも膜下出血∗; 心臓弁膜症∗; 毎日を楽しんでいる; 性別; 体力の低下を感じる; 突然の言語障害; 運動; 血縁者に脳卒中; 健康診断を受けている; 心拡大; 栄養バランス.

スコア計算対象外の質問: 身長; 体重; 喫煙開始年齢; 禁煙開始年齢; 健康について気になること.

(2003 年 4 月 11 日受付)
(2003 年 6 月 11 日採択)