文書分類技法とそのアンケート分析への応用

A Document Classification Method and its Application to Questionnaire Analyses

平澤 茂一† 石田 崇‡ 足立 鉱史‡§ 後藤 正幸* 酒井 哲也**

Shigeichi Hirasawa† Takashi Ishida‡ Hiroshi Adachi‡ Masayuki Gotoh* Tetsuya Sakai**

† 早稲田大学理工学部 経営システム工学科
‡ 早稲田大学 大学院理工学研究科 (§ 現 新日鉄ソリューションズ (株))
* 武蔵工業大学環境情報学部, **(株) 東芝研究開発センター

† School of Science and Engineering, Waseda University
‡ Graduate School of Science and Engineering, Waseda University
* Faculty of Environment and Information Studies, Musashi Institute of Technology
** Knowledge Media Labolatory, Research and Development Center, Toshiba Co., Ltd.

要旨： 情報検索技術をベースとする文書分類と文書クラスタリング技法とその応用について報告する．まず，確率的潜在意味インデキシングモデルを用いた文書分類技法が良い分類性能を持つことをベンチマークデータを用いて示す．次に，これを文書クラスタリング技法に拡張し，定型型の回答項目と自由記述型の回答項目の混在する学生アンケートの回答文に適用する．その結果，授業運営・教務事務改善に役立つ情報が抽出できることを示す．

Abstract: Using information retrieval techniques especially based upon Probabilistic Latent Semantic Indexing (PLSI) model, we discuss on methods for classification and clustering of documents. First, a method for classifying documents with free format is proposed. The effectiveness of the method is demonstrated by using the benchmark test set in Japanese, called BMIR-J2. Second, the clustering method is developed by modifying the classification method and is applied to documents obtained from student questionnaires, where the questionnaires are composed of two types of questions, i.e., answered with fixed format and with free format. The results obtained are evaluated as categorical documents. Finally, the documents are also partitioned into clusters and analyses of questionnaires are executed from the standpoint of statistical techniques to documents with fixed format. Statistical interpretation is added to the results and new knowledge obtained by it gives us effective facts for classroom management and leads to faculty developments.

1 はじめに

WWW 検索などの情報検索技術はコンピューティングパワーの飛躍的な増大，メモリの格段の低コスト化・大容量化により 1990 年代後半，本格的に実用化され今日に至っている．また，文書分類・文書クラスタリング技術も多くは文書検索技術に基づいて開発されている．

一方，大量の電子化された文書データベースから，与えられた文書と類似した文書を抽出したり，隠された知識を発見しようとするテキストマイニングのニーズが高まっている．性能の良いテキストマイニング手法のニーズは大学教育の現場にもある．例えば，日本技術者認定機構 (JABEE) が実施する専門分野を単位とする認定プロセスがある．ここでは授業実施の後，学生の授業評価による恒常的なフィードバックループによる教員の改善努力・授業管理が要求される．筆者らはここ 3 年間，「コンピュータ工学」と「情報化社会概論」の 2 科目について授業改善のための学生アンケートを実施した．学生の成績や授業満足度を説明するための授業モデルを提案し，授業改善のためのアンケート分析結果を報告した [1]-[9].

本稿では，情報検索技術をベースとした文書分類・クラスタリング技法に焦点をあて，性能評価とその応用について考察する．まず，情報検索モデルとしては確率モデルに分類される確率的潜在意味インデキシング (Probabilistic Latent Semantic Indexiing:PLSI) モデル [10] に基づく文書分類技法 [4] を提案する．この分類技法は PLSI モデルで実行する EM アルゴリズムが初期値に依存するという性質を巧みに利用している．次に，これをベンチマークデータ BMIR-J2[11] に適用し，分類性能を明らかにする．その結果，比較的小規模のデータベースに対し優れた性能を持つことを示す．さらに，定型型項目（選択式）と自由記述型項目（記述式）の混在する文書をも対象とする文書クラスタリング技法 [12][13] に拡張し，学生アンケートに適用する．その結果，異なる科目を履修する学生に同一のアンケートを実施したとき，クラスタリング誤り（科目をカテゴリと見た分類誤り）が他の情報検索モデルに基づく技法と比べ小さくできることを明らかにする[1]．次に，授業評価に関する学生アンケートに応用し，クラスタリング結果を統計処理することにより有効な知識が発見出来ることを示す．同時に，「WEB を用いた科目登録システム」の評価に関するアンケートに適用した結果についても述べる．なお，技法の数理的考察は稿を改める．

[1] ここではこの問題をクラス併合・分類問題と呼ぶ．

2 情報検索モデル

情報検索システムの数学的モデル [14] は大別して 2.1 ～ 2.3 で述べる 3 つである．ここではさらに 2.4 と 2.5 に，本研究のベースラインシステムである潜在意味インデキシング (Latent Semantic Indexing:LSI) モデルと PLSI モデルを示す．

文書を $d_j(j = 1, 2, \cdots, D)$，文書集合に現れる単語を $t_i(i = 1, 2, \cdots, T)$ とする．多くの検索モデルは，文書集合を表現する単語－文書行列 $A = [a_{ij}]$ と，検索質問 q の行列 A に対応する質問ベクトル $\boldsymbol{q} = (q_1, q_2, \cdots, q_T)^{\mathrm{T}}$ により特徴付けられる．$A = [a_{ij}]$ は一般に文書 d_j における単語 t_i の重要度を表す．さらに，文書と検索質問間の類似度 $s(q, d_j)$ が定義される．類似度は q に対する d_j の適合の度合いを示すから，ランキング出力が可能である．ここで，文書 d_j を文書ベクトル $\boldsymbol{d_j} = (a_{1j}, a_{2j}, \cdots, a_{Tj})^{\mathrm{T}}$ で表す．

2.1 ブーリアンモデル

本モデルにおける単語は特に文書に付けられたキーワードである．したがって

$$a_{ij} = \begin{cases} 0, & t_i \not\subset d_j \\ 1, & t_i \subset d_j \end{cases} \tag{2.1}$$

$$q_i = \begin{cases} 0, & t_i \not\subset q \\ 1, & t_i \subset q \end{cases} \tag{2.2}$$

である．ただし，$t \subset d$ は単語（キーワード）t が文書 d に現れることを示す．質問ベクトル $\boldsymbol{q} \supset \boldsymbol{d_j}$ ($\boldsymbol{q} \wedge \boldsymbol{d_j} = \boldsymbol{q}$ のとき）ならば文書 d_j は適合と判定する．通常，検索質問 q には論理和・否定を含むため，q を積和標準形に展開しそれぞれの項に対し適合判定を行う．

2.2 ベクトル空間モデル (Vector Space Model:VSM)

ブーリアンモデルに部分照合機能を持たせたモデルである．そのため，一般に $a_{ij} \geq 0$，$q_i \geq 0$ は非 2 値で，それぞれ文書 d_j，質問 q に出現した単語 t_i の重み（重要度）を用いる．多くの場合，a_{ij} は文書 d_j における単語 t_i の出現頻度 (trem frequency:tf) と全文書中で単語 t_i の出現した文書数の逆数 (inverse document frequency:idf) の積 (tf-idf 値) を用いる．

前者は局所的重み係数，後者は大域的重み係数である．文書 d_j と検索質問 q の間の類似度は，通常それぞれのベクトル $\boldsymbol{d}_j$，$\boldsymbol{q}$ の余弦（次式）を用いる．

$$s(q, d_j) = \boldsymbol{q}^{\mathrm{T}} \boldsymbol{d}_j / |\boldsymbol{q}||\boldsymbol{d}_j| \quad (0 \leq s(\ ,\ ) \leq 1) \tag{2.3}$$

2.3 確率モデル

検索質問 q が与えられたとき，真に適合する文書のみを含む集合 $\mathcal{D}^*$（真の適合集合）が存在する．正確にはその属性を知らないが，この属性を特徴付ける単語の集合があることは分かっているとする．そこで，q に対する検索結果から，それぞれの文書の適合・不適合を決め，これを繰り返して次第に真の適合集合 $\mathcal{D}^*$ を見つけ出す過程を確率的手法によりモデル化する．ここで，単語－文書行列 A，質問ベクトル $\boldsymbol{q}$ は共に 2 値で，ブーリアンモデルと同様にそれぞれ式 (2.1),(2.2) で与えられる．

2.4 LSI モデル

VSM における単語－文書行列 A を特異値分解 (SVD) して得られた行列 A_K を用いて検索を行う．SVD はノイズを除去し，検索性能を向上させることが知られている．ここで，$A \in \mathcal{R}^{T \times D}$ のとき

$$A \to A_K = U_K \Sigma_K V_K^{\mathrm{T}} \tag{2.4}$$

とすれば，$U_K \in \mathcal{R}^{D \times K}$, $\Sigma_K \in \mathcal{R}^{K \times K}$, $V_K \in \mathcal{R}^{T \times K}$, $K \leq p \leq \max\{T, D\}$ である．ただし，p は行列 A のランクである．文書ベクトル $\boldsymbol{d}_j$ は行列 Σ_K により $\dot{\boldsymbol{d}}_j = \Sigma_K V_K^{\mathrm{T}} \boldsymbol{e}_j \in \mathcal{R}^{K \times 1}$ に変換される．質問ベクトル $\boldsymbol{q} \in R^{T \times 1}$ も $\dot{\boldsymbol{q}} = \Sigma_K^{-1} \in R^{K \times 1}$ に変換され，このとき類似度は式 (2.3) と同様，$\dot{\boldsymbol{d}}_j, \dot{\boldsymbol{q}}$ を用いて与えられる．ただし，$\boldsymbol{e}_j = (e_1, e_2, \cdots, e_T)^{\mathrm{T}}$ のとき $e_j = 1$, $e_{j'} = 0 (j' \neq j)$ である．

2.5 PLSI モデル [10]

LSI モデルが特異値分解という行列の代数的圧縮を行っているとすれば，PLSI モデルは隠れ状態（潜在クラス）による行列の確率的圧縮を行っているといえる．すなわち，K 個の隠れ状態 $z_1, z_2, \cdots, z_K$ $(K \leq \max\{T, D\})$ により分解され，これを合成したモデルである．ここで，(1) 組 (t_i, d_j) は互いに独立，(2) 状態 z_k の下に t_i,d_j は（条件付）独立，を仮定する．その結果，次式の対数尤度関数

$$L = \Sigma_{i,j} a_{ij} \log \Pr(t_i, d_j) \tag{2.5}$$

を最大化する EM アルゴリズムを用いて $\Pr(d_j)$, $\Pr(t_i|z_k)$, $\Pr(z_k|d_j)$ が求まる．類似度は次式を用いて，式 (2.3) により計算する．

$$\boldsymbol{q} = (\Pr(q|z_1), \Pr(q|z_2), \cdots, \Pr(q|z_K))^{\mathrm{T}} \in \mathcal{R}^{K \times 1} \tag{2.6}$$
$$\boldsymbol{d}_j = (\Pr(d_j|z_1), \Pr(d_j|z_2), \cdots, \Pr(d_j|z_K))^{\mathrm{T}} \in \mathcal{R}^{K \times 1} \tag{2.7}$$

3 文書分類・クラスタリング技法の提案

文書を S 個のカテゴリ $C_1, C_2, \cdots, C_S$ に分類する問題を考える．既に分類された D_L 個の文書（学習データ）に対応する文書ベクトル $\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_{D_L}$，これから分類しようとする D_T 個の文書（評価データ）に対応する文書ベクトルを与えるものとする．

3.1 従来の分類法

従来の検索モデルでは，いずれも類似度の計算方法が与えられている[2]．したがって，従来の検索モデル (VSM, LSI モデル，PLSI モデル) に基づく分類方法は類似度を用いて，次のように与えることが出来る．

[従来の分類アルゴリズム]

(1) 学習文書からカテゴリ毎の重心ベクトル $\overline{\boldsymbol{d}}_s$ $(s = 1, 2, \cdots, S)$ を求める．

(2) 評価文書 d_j を $s(\overline{\boldsymbol{d}}_s, \boldsymbol{d}_j)(s = 1, 2, \cdots, S)$ が最大となるカテゴリ $C_{\hat{s}}$ に分類する． □

ここで，PLSI モデルにおいては式 (2.7) で与えたように，行列 $[\Pr(d_j|z_k)]$ 上で計算する．

3.2 提案分類法 [4][13]

PLSI モデルで $S = K$ とし，EM アルゴリズムの初期値依存性を利用する．ここで，カテゴリは 1 つの隠れ状態（潜在的クラス）に対応すると考える．

[PLSI モデルを用いた分類アルゴリズム]

(1) 学習文書からカテゴリ毎の重心ベクトル $\overline{\boldsymbol{d}}_s \in \mathcal{R}^{T*1} (s = 1, 2, \cdots, S)$ を求める[3]．

(2) 重心ベクトル $\overline{\boldsymbol{d}}_s$ を $z_k (k = s)$ に対する初期値として EM アルゴリズムを実行し，$\Pr(t_i|z_k)$,$\Pr(d_j|z_k)$,$\Pr(z_k)$ を求める[4]．

(3) 評価文書 d_j を $\Pr(z_k|d_j)(k = 1, 2, \cdots, K)$ が最大となるカテゴリ $C_{\hat{k}}$ に分類する． □

3.3 提案クラスタリング法 [12]

PLSI モデルをクラスタリング法に拡張する．ここでは，クラスタ数を S とし，$S \leq K$ に選ぶ．

[PLSI モデルを用いたクラスタリングアルゴリズム]

(1) Tempered EM アルゴリズムを用いて，$\Pr(t_i|z_k), \Pr(d_j|z_k), \Pr(z_k)$ を求める．

(2) 文書 d_j を $\Pr(z_k|d_j)(k = 1, 2, \cdots, K)$ が最大となるクラスタ $c_{\hat{k}}$ に分類．もし，$S = K$ ならば終了．

(3) もし，$S < K$ ならば

$$s(z_k, z_{k'}) = \boldsymbol{z}_k^{\mathrm{T}} \boldsymbol{z}_{k'} / |\boldsymbol{z}_k||\boldsymbol{z}_{k'}| \tag{3.1}$$
$$\boldsymbol{z}_k = (\Pr(t_1|z_k), \Pr(t_2|z_k), \cdots, \Pr(t_i|z_k))^{\mathrm{T}} \tag{3.2}$$

を計算し，類似度 $s(z_k, z_{k'})$ を距離測度とする群平均法によりクラスタ数が S になるまでクラスタ凝集を実行する． □

3.4 選択式・記述式からなる文書

選択式・記述式のフォーマットが混在する文書が扱えるよう拡張する．その単語－文書行列 A の作成法は合成確率モデル [15] を修正し，次のように与える [12].

$$A = \begin{bmatrix} \lambda G \\ (1-\lambda) H \end{bmatrix} \quad (0 \leq \lambda \leq 1) \tag{3.3}$$

ただし，$G \in \{0, 1\}^{I*D}$, $H \in \mathcal{R}^{T*D}$ である．前者は選択式回答から求めたアイテム－文書行列，後者は記述式回答から求めた単語－文書行列である．

4 文書分類技法の実験と性能評価

ここでは，表 4.1 に示す文書集合 (a)(b) を対象に（分類）実験を行う．

---

[2] ブーリアンモデルについても類似度を定義することが出来る．ただし，$s(\ ,\ ) = \{0, 1\}$ である．

[3] 重心ベクトル $\overline{\boldsymbol{d}}_s$ の代わりに，カテゴリ C_s の典型的ベクトル $\boldsymbol{d}_s^*$ を用いても良い．

[4] 初期値を与える際，一般に $\Sigma_i \Pr(t_i|z_k) = 0$ となる z_k があり，ラプラス推定量を用いるなどの工夫が必要である．

表 4.1: 対象文書集合

内容	内容	文書数 D	単語数 T	カテゴリ数 S
(a) '94 年毎日新聞記事	記述式	101,058	107,835	9+1
(b) 学生アンケート	選択式＋記述式	170	3,293+3,993	2
(c) 学生アンケート	選択式＋記述式	111	3,293+3,993	
(d) 学生アンケート	選択式＋記述式	3,819	2,970	

4.1 ベンチマークデータ BMIR-J2 の分類

表 4.1(a) は情報処理学会が用意した文書検索・文書分類問題のためのベンチマークデータ BMIR-J2[11] である．カテゴリは経済，政治，スポーツ，読書，科学，社会，生活など 10 個からなる．

[実験 1-1] { 経済，スポーツ，社会 } のそれぞれのカテゴリから，ランダムに評価文書 $D_T = 50$ 個を取り出す．学習文書数 $D_L = 50$ で同数である．分類誤り率 P_c に関する評価結果を表 4.2 に示す．

表 4.2: 各分類法の分類誤り率

分類法	分類誤り率 P_c (%)	備考
VSM 法	42.7	
LSI 法	38.7	$K = 81$（累積寄与率 70 %）
PLSI 法	20.7	$K = 10$
提案法	6.0	$K = 3$

[実験 1-2]

カテゴリを S 個選ぶ．それぞれのカテゴリから評価文書 D_T 個を取り出す．ただし，学習文書数 D_T は評価文書数と同数 ($D_L = D_T$) である．$D_L = D_T = 50 \sim 450$, $S = 2 \sim 8$ に対する分類誤り率 P_c に関する評価結果を図 4.1,4.2 に示す．

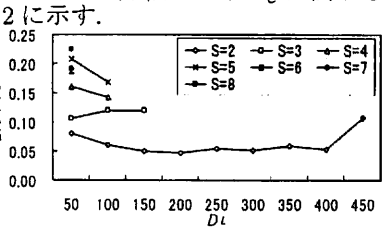

図 4.1: 学習文書数 D_L に対する分類誤り率 P_c

図 4.1,4.2 より，学習文書数 D_L，カテゴリ数 S の増加にしたがい分類誤り率 P_c が増加することが分かる．

4.2 授業に関する学生アンケートの分類

表 4.1(b) は理工学部経営システム工学科設置 2 年前期・必修科目「コンピュータ工学」とメディアネットワークセンター設置全学共通・選択科目「情報化社会概論」の授業に関する学生アンケートである．両方の科目履修者に初回 (IQ) と最終回 (FQ) 同一のアンケートを実施している．表 4.3 に主な質問内容を示す．

[実験 2] 「コンピュータ工学」の履修学生 135 名と「情報化社会概論」の履修学生 35 名のアンケートシート（文書）を併合し，提案クラスタリング法でクラスタリングした．これを 2 クラスの併合・分類問題として評価した結果を図 4.3,4.4 に示す．ここで，アンケートは選択式と記述式からなるので，3.5 で述べた拡張した単語ー文書行列を用いている．この実験は正確にはクラスタリング問題であるが，両科目の履修学生は一方が理系，他方が文系と明らかに履修学生特性が異なるため，履修科目をカテゴリとみなして分類誤り率 P_c で評価した．その結果，次のことが分かる．

(1) 図 4.3 より，$\lambda = 0.5, K = 5$ のとき，最も小さい分類誤り率が得られる．また，VSM を用いた結果より大きく改善されている．

(2) 図 4.4 より，$K = 5 \sim 10$ を選ぶと分類誤り率は小さくなる．

この他，$\lambda = 0.5, K = 2$ のとき，対数尤度は最も小さく，K が増加するにしたがい単調に対数尤度は大きくなることが分かっている．

以上から，λ に最適値があり，対数尤度 L と K の選び方により分類誤り率 P_c を小さく出来ることが分かる．

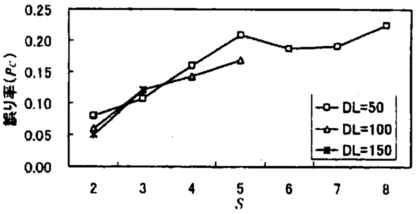

図 4.2: カテゴリ数 S に対する分類誤り率 P_c

表 4.3: 初回アンケート (IQ) 内容の例

項目	例（小質問）
選択式	・コンピュータの利用経験年数はどの位か ・海外留学の計画を持っているか ・PC を組み立てられるか ・情報関係の資格を持っているか ・知っている情報関係の術語を 10 個書きなさい
記述式	・コンピュータの知識・経験について書きなさい ・卒業後どのような仕事につきたいか ・科目名からどのような講義内容をイメージするか ・講義内容は理解できたか

5 学生アンケート分析

ここでは，表 4.1 に示す文書集合 (c)(d) を対象に（分類・クラスタリング）実験を行う．

5.1 授業に関する学生アンケートの分類 [7]

初回アンケート (IQ) のみから「コンピュータ工学」のクラスを 2 分割する問題（クラス分割問題）を考える．表 5.1 のように，講義内容による 2 つのクラスを設定する．

表 5.1: クラス S・クラス G の講義内容

クラス	講義内容
S	・アーキテクチャ全般（スタック，プロセッサ，メモリなど） ・ハードウェア全般（論理設計，論理回路，オートマトンなど） ・ソフトウェア全般（オペレーティングシステム，UNIX など）
G	・コンピュータの歴史，基本構成 ・コンピュータアーキテクチャ基礎 ・ハードウェア基礎 ・ソフトウェア基礎 ・周辺技術（情報通信，ネットワーク，人工知能など）

[実験 3] 3.2 で述べた提案分類法の初期値（典型的ベクトル）として，既に将来の進路（就職先）の決まった上級生によるアンケートを用いた．その結果，履修学生 135 名の内，自分の希望するクラスを回答した 111 名のアンケーを合わせ，表 5.2 のように分類された．この表にしたがい，それぞれのクラスに属する学生の選択式回答の判別分析結果を図 5.1 に示す．

5.2 WEB 科目登録に関する学生アンケートのクラスタリング

本年 3,4 月に全学的な WEB を用いた科目登録を行った [16]．これに先立ち，昨年度は学部を限定して実施した．その際，主として利用性を評価するために，昨年 5 月「WEB 科目登録システム」のアンケート調査を行った．表 5.3 に主な質問内容を示す．

[実験 4] S=2,5 について，$0.0 \leq \lambda \leq 1.0$ と変化させる．$\lambda = 0.3$ のとき，$S = 2$ のクラスタと $S = 5$ のクラスタが図 5.2 のように分割出来る．また，それぞれのクラスタから特徴単語（選択式回答を含む）抽出を行った結果を図 5.3 に示す．

6 考察

第 5 章で行った実験について考察する．

6.1 授業アンケートからのクラス分割

クラス分割は IQ から学生の潜在的特性を把握することによって，可能な限り各学生に適したクラスを提供することが目的である．学生の選択を単純に正解と仮定すると分類誤り率は大きい．しかし

(1) 客観的に明確な差異を持つ 2 つのクラスへの分類問題ではない．

(2) クラス分割結果を学生指導に用いることができる．

と考えられる．また，

(3) 結果の検証には追跡調査が必要である．

このようなクラス分割は講義内容によるものだけではなく，(a) 学生のレベルによる分割，(b) 運営方法によるクラス分割なども検討が必要である [8][9]．

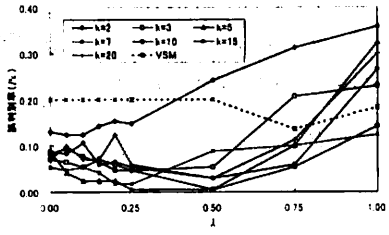

図 4.3: λ に対する分類誤り率 P_c

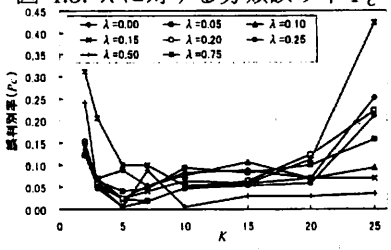

図 4.4: 隠れ状態数 K に対する分類誤り率 P_c

6.2 WEB 科目登録アンケートからのクラスタリング

このアンケートの目的は学生がこのシステムをどう見ているか，改善するとすればそれはどこかについてヒントを得ることである[5].

(1) $c_5(2)$ は実際に WEB 科目登録を実行した大半の学生が含まれている．その内，利用に（問題なしと答えた）肯定的な学生の割合は 541/581 ≒ 0.933 である．

(2) 自宅から WEB 科目登録を行った学生の大半は $C_2(1) \wedge C_5(2)$ に含まれる．

今後実際の登録データを参照し，学生の所属学部・登録内容を含む選択式回答を統計処理し，特徴を詳細に抽出する必要がある．

7 むすび

ベンチマークデータや学生アンケートを用いて，提案した PLSI モデルに基づく分類法・クラスタリング法が有効であることを示した．特に，文書集合の規模が比較的小さい場合良い性能を持つ．選択式と記述式の文書にも使えるよう拡張し，応用分野も広い．

分類・クラスタリング技法はアンケート分析方法において，その手法の一つに過ぎない．統計的処理や特徴文・重要文抽出手法 [1] と合わせ，総合的に分析する必要がある [7]-[9]．また，アンケートの設計には対象とするシステムのモデルを作り，その入力としてアンケートの質問項目を設計しなければならない [2][3][5]．その結果，入力と出力の関係を明らかにしてモデルの説明が出来，ここから新しい知識が得られる．目的に合わせアンケートを設計しなければ，望む分析結果は得られない．既存のアンケートから新しい知識を発見することは難しいのである．

謝辞：アンケート分析にご協力を戴いた武蔵工業大学環境情報学部 後藤研究室 渡辺智幸君に感謝します．本研究の一部は早稲田大学特定研究課題 2004A-174 の助成による．

参考文献

[1] 酒井哲也，伊藤潤，後藤正幸，石田崇，平澤茂一，“情報検索技術を用いた効率的な授業アンケートの分析，”2003 年経営情報学会全国春季研究発表大会予稿集，pp.182-185，東京，2003 年 6 月．

[2] 後藤正幸，酒井哲也，伊藤潤，石田崇，平澤茂一，“選択式・記述式アンケートからの知識発見，”PC カンファレンス予稿集，pp.43-46，鹿児島，2003 年 8 月．

[3] 平澤茂一，石田崇，伊藤潤，後藤正幸，酒井哲也，“授業に関する選択式・記述式アンケートの分析，”平成 15 年度大学情報化全国大会．pp.145-145，東京，2003 年 9 月．

[4] 伊藤潤，石田崇，後藤正幸，酒井哲也，平澤茂一，“PLSI を利用した文書からの知識発見，”2003 年 FIT 論文集，pp.83-84，2003 年 9 月．

表 5.2: クラス S・G の分類結果

クラスタリング	学生自身の選択		
	G	S	合計
G	29	20	49
S	34	28	62
合計	63	48	111

[5]WEB 科目登録を実行した割合は 755/3855 ≒ 0.199 である．

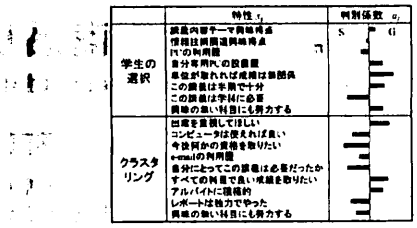

図 5.1: クラス S・G の属性の判別分析結果

表 5.3: アンケート内容の例

選択式	科目登録は完了したか WEB により科目登録したか マークシートに比べ WEB の利点は何か どのようなネットワーク環境でアクセスしたか 申請画面は分かりやすかったか 登録されている科目一覧の印刷機能を使ったか FAQ を参照したか 登録後送信する科目登録結果のメールは必要か
記述式	WEB を使わなかった人の理由は何か WEB 科目登録に問題があった人の理由は何か 申請方法が分かりにくかった人の理由は何か 「科目登録情報ページ」が分かりにくかった人の理由は何か

[5] 石田崇，伊藤潤，後藤正幸，酒井哲也，平澤茂一，“授業モデルとその検証，”2003 年経営情報学会全国秋季研究発表大会予稿集，pp.226-229，函館，2003 年 11 月．

[6] 酒井哲也，石田崇，後藤正幸，平澤茂一，“自然言語表現に基づく学生アンケート分析システム，”2004 年 FIT 論文集，pp.325-328，2004 年 9 月．

[7] S. Hirasawa, “Knowledge discovery from documents–A case of improvements for quality of education–,” A short course at Tamkang University, Taipei, R.O.C., May 2004.

[8] 石田崇，足立紘史，後藤正幸，酒井勝也，平澤茂一，“情報検索技術を用いた選択式・自由記述式の学生アンケート解析，”経営情報学会 2004 年度秋季全国研究発表大会予稿集，pp.466-469，名古屋，2004 年 11 月．

[9] 石田崇，後藤正幸，平澤茂一，“大学の情報系授業における学生アンケートの分析，”コンピュータ・エデュケーション誌，vol.18，2005（掲載予定）．

[10] T. Hofmann, “Probabilistic latent semantic indexing,” Proc. of SIGIR'99, ACM Press, pp.50-57, 1999.

[11] 毎日新聞社，CD．毎日新聞 '94，日外アソシエーツ，1995 年．

[12] S. Hirasawa, and W. W. Chu, “Knowledge acquisition from documents with both fixed and free formats,” *Proc. IEEE 2003 Int. Conf. on SMC*, pp.4694-4699, Washington DC, Oct. 2003.

[13] S. Hirasawa, and W. W. Chu, “Classification methods for documents with both fixed and free formats by PLSI model,” *Proc. 2004 International Conference on Management Science and Decision Making*, Taipei, Taiwan, May 2004.

[14] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, New York, Addison-Wesley, 1999.

[15] D. Cohn, and T. Hofmann, “The missing link–A Probabilistic model of document content and hypertext connectivity,” *Advances in Neural Information Processing Systems (NIPS*13)*, MIT Press, 2001.

[16] 久保田学，“履修科目申請システム成功への過程と OSS 活用への取り組み，”*Linux World and Expo*，東京，2005 年 6 月（発表予定）．

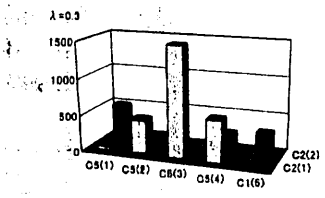

図 5.2: $S = 2$ と $S = 5$ のクラスタ

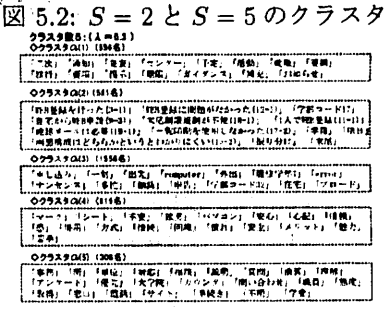

図 5.3: クラスタからの特徴単語抽出結果（$S = 2$）