PLSIによる学生アンケートからの知識発見
− 日台学生アンケート分析 −
Knowledge Discovery from Student Questionnaire based on PLSI

長尾 壮史† 坂口 朋章† 石田 崇‡ 平澤 茂一‡
Masafumi NAGAO† Tomoaki SAKAGUCHI† Takashi ISHIDA‡ Shigeichi HIRASAWA‡
† 早稲田大学 大学院理工学研究科経営システム工学専攻
‡ 早稲田大学 理工学部経営システム工学科
† Graduate School of Science and Engineering, Waseda University
‡ School of Science and Engineering, Waseda University

要旨： 大学における授業の改善を最終目標とした学生アンケートの分析手法および分析結果について報告する．アンケート回答の分析は人の主観に頼るなど適切な手法が確立しているとは言えず，現在も盛んに研究が進められている．本研究では，情報検索分野でその有効性が示されている PLSI (Probabilistic Latent Semantic Indexing) を用いることにより，選択式回答と記述式回答の混在したアンケートを統合的に分析する手法を示す．また，この手法を用いて日本と台湾の両大学で行われたアンケートを分析しその結果の比較を行う．

Abstract: This paper analyzes student questionnaires for improvements in quality of education. A clustering method based on PLSI model is presented and is applied to the analysis of the questionnaires which are conducted at the University in both Japan and Taiwan. Then we analyze the results of the clustering and show that we can grasp the tendency of students.

1 はじめに

大学における授業改善の取り組みの一環として学生アンケートが広く行われている．多くの場合，アンケートは選択式回答項目と自由記述式の回答項目から構成されている．しかし，選択式回答に対しては頻度の集計や評価値の平均・分散を求めるにとどまり，また，自由記述式回答に対しては，学生の自由な意見を集めることができるものの，実際には人が読んだ上で主観的な判断により集計されているのが現状である．また，一般的には選択式と記述式の回答はそれぞれが別々に分析され，統合的に分析する適切な処理方法は未だ確立しているとは言えない．

一方，情報検索技術の分野では近年の情報技術の発展に伴って新たな進展を見せており，文書を扱う様々な手法が提案されている．代表的な手法の 1 つとして確率空間を利用して文書索引語行列を低次元に圧縮する PLSI (Probabilistic Latent Semantic indexing) [1] が提案され，情報検索におけるその有効性が示されている．

本研究では，PLSI を利用することで選択式・自由記述式アンケートを統合して効率的に分析する．PLSI によりアンケート分析を行う手法は既に提案されており [2]，これを日本と台湾の大学それぞれで実施されたアンケート回答に適用する．この手法を用いると，概念的に類似した学生同士がクラスタリングされるため，学生全体がある傾向を持ついくつかのグループに分割される．さらに，PLSI を用いた特徴語の特徴文抽出手法 [3] を利用し，各グループの特性を探ることによって授業改善に有用な知見が得られることを示す．

2 潜在的意味モデル

2.1 LSI モデル

S. Deerwester らは，意味的情報検索のモデルとして LSI (Latent Semantic Indexing) [4] を提案した．LSI では，文書−索引語行列 A を特異値分解 (SVD) によって $A = U\Sigma V^{\mathrm{T}}$ と分解する．このうち，固有値の大きい方から K 個を用いて $A_K = U_K \Sigma_K V_K^{\mathrm{T}}$ とすることにより，A を K 次元の潜在的意味空間に圧縮することでノイズの除去を行う．これは行列 A と A_K の 2 乗最小誤差を最小とする圧縮となっている．しかし，LSI では理論的に適切な重み付け方法などの問題がある．

2.2 PLSI モデル

一方，T. Hofmann によって提案された PLSI [1] は，行列の代数的圧縮である LSI と異なり，確率モデルに基づいて圧縮を行う手法である．

PLSI では，意味的な隠れ属性 z_k $(k = 1, 2, \cdots, K)$ のもとで文書 d_i $(i = 1, 2, \cdots, N)$ と索引語 w_j $(j = 1, 2, \cdots, M)$ の生起は独立であるとし，d_i と w_j の同時確率を

$$P(d_i, w_j) = \sum_k P(z)P(d_i|z_k)P(w_j|z_k) \quad (1)$$

として与えるモデルである．ここで，文書 d_i における索引語 w_j の実際の出現回数を $n(d_i, w_j)$ とすると，次式の対数尤度

$$L = \sum_{i,j} n(d_i, w_j) \log P(t_i, d_j) \quad (2)$$

を最大化するような $P(z_k)$, $P(d_i|z_k)$, $P(w_j|z_k)$ を EM アルゴリズムを用いて推定する．

[E-step]

$$P(z_k|d_i, w_j) = \frac{P(z)P(d_i|z_k)P(w_j|z_k)}{\sum_{k'} P(z_{k'})P(d_i|z_{k'})P(w_j|z_{k'})} \quad (3)$$

[M-step]

$$P(d_i|z_k) = \frac{\sum_i n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{i',j} n(d_{i'}, w_j)P(z_k|d_{i'}, w_j)} \quad (4)$$

$$P(w_j|z_k) = \frac{\sum_i n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{i,j'} n(d_i, w_{j'})P(z_k|d_i, w_{j'})} \quad (5)$$

$$P(z_k) = \frac{\sum_{i,j} n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{i,j} n(d_i, w_j)} \quad (6)$$

収束するまで，E-step と M-step の計算を繰り返すが，実際には過学習を避けるため Temperd EM を用いている [1].

3 PLSI によるクラスタリングと特徴語抽出

3.1 文書クラスタリング

PLSI における隠れ属性 z_k $(k = 1, 2, \cdots K)$ はひとつの概念を表していると捉えることができるため，PLSI モデルを用いて対象文書をクラスタリングすることができる [2]. いま，文書集合を S 個のクラスタにクラスタリングするとする．ただし $S \leq K$ である．

[PLSI によるクラスタリングアルゴリズム]

1. PLSI を実行し，$P(z_k), P(d_i|z_k), P(w_j|z_k)$ を求める．
2. 各文書 d_i を，$P(z_k|d_i)$ が最大となるような z_k に割り当てる．
3. $K = S$ ならば，各 z_k がそれぞれのクラスタとなる．$K > S$ の場合，各 z_k 同士の類似度を測り，距離尺度として $K = S$ となるまで群平均法でクラスタ併合を行う．

3.2 選択式・自由記述式アンケートのクラスタリング

PLSI によるクラスタリング手法を拡張することにより，選択式・自由記述式が混在した文書に対して適用することが可能である [2].

文書–索引語行列 A を，選択式・自由記述式の合成行列として以下のように拡張して，PLSI によるクラスタリングを実行する．

$$A = \begin{bmatrix} \lambda G \\ (1-\lambda)H \end{bmatrix}, \quad (0 \leq \lambda \leq 1) \qquad (7)$$

ここで，選択式の行列 $G \in \{0,1\}^{I*N}$，自由記述式の行列 $H \in R^{M*N}$ である．ただし，I は選択式回答の項目数とする．また，λ は選択式・自由記述式を合成する際の重みを表している．

3.3 特徴語・特徴文抽出法 [3]

ここで，PLSI における $P(w_j|z_k) - P(w_j)$ を各単語のスコアとし，文を構成する単語スコアの総和を文のスコアとする．各クラスタにおいてスコアの高い文を特徴文とする．

4 学生アンケート

アンケートは授業の初回 (IQ) と最終回 (FQ) の二回実施した．アンケートを行った対象は，「コンピュータ工学」[5] (早稲田大学理工学部経営システム工学科 2 年生必修科目) 履修生 (理系) および「情報化社会概論」(早稲田大学メディアネットワークセンター設置科目) 履修生 (文系) である．また同じアンケートを中国語に翻訳し，台湾の大学で同様の科目においても実施した．表 1 にアンケート内容の例を示す．

表 1: アンケート項目の一例

初回アンケート (IQ)：

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． ・明確な目的を持ってこの講義に臨んでいるか． ・情報関係の資格を持っているか． ・この講義は自分にとって必要か．
記述式	・コンピュータの知識・経験について書きなさい． ・卒業後どんな仕事に就きたいか． ・科目名からどのような講義内容をイメージするか．

選択式項目：7 質問 (各 4-20 小質問)
記述式項目：5 質問 (各 250-300 文字)

5 学生アンケート分析結果

大勢の学生に対して画一的な講義を行うよりも，学生の志向やレベルなど特性に応じた講義を実施することが望ましいといえる．そこで，IQ から学生の潜在的な特性を把握することにより授業クラスをいくつかに分割する問題を考える．すなわち，IQ の回答に対して PLSI クラスタリングを行い，概念的に類似した学生同士を複数のクラスに分割する．講義の開始時に IQ によって学生が潜在的に希望する講義内容を予測して学生をクラス分けすることができれば，学生側と教員側にとって満足度の高い講義が可能になると考えられる．

以下では，

- 日本の学生だけからなる集合に対するクラスタリング
- 日本と台湾の全学生からなる集合に対するクラスタリング

の結果について報告する．

5.1 日本の学生アンケート

「コンピュータ工学 (CE)」履修生 (理系学生：125 人) と「情報化社会概論 (IS)」履修生 (文系学生：19 人) の IQ をマージしたデータに対して PLSI によるクラスタリングを行う．理系と文系という特性はもっとも傾向の違いが顕著であると考えられる．実際に，マージした IQ をクラスタリングによって 2 クラスに分割した結果，理系か文系かという属性が大きく影響していることが判明した．そこで，ここでは学生の属性 (理系 or 文系) をカテゴリと考えた学生の分類問題と捉えることとする．

図 1 に隠れ属性数 K を変化させてクラスタリング $(S = 2)$ を行ったときの結果を示す．なお，誤分類率とは CE の学生を文系のクラスへ，IS の学生を理系のクラスへ誤って分けてしまった誤り確率である．

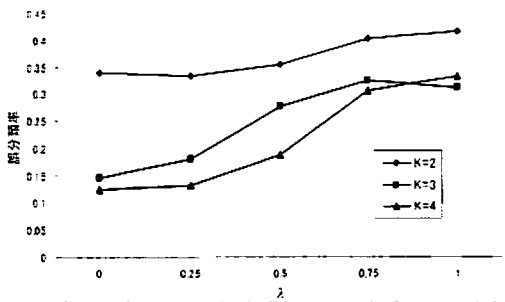

図 1: 隠れ属性数 K を変化させたときの誤分類率

5.2 日本の学生アンケートに関する結果と考察

5.2.1 文系・理系のクラス分割問題

(1) 隠れ属性数 $K = 2$ の場合

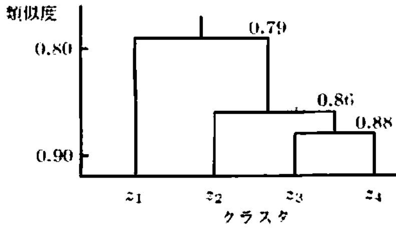

図 2: $K=4$ のときのデンドログラム

誤分類率

図 1 より，隠れ属性数 K が小さいほど誤分類率が高くなることが分かる．これは，経営システム工学科（CE 履修生）という学科のもつ特性が理由のひとつであると考えられる．本学科の学生は卒業（または修了）後の就職先が多岐に渡り，理工学部の他学科と比べても比較的文系的志向の強い学科であるといえる．実際に例年の進路先は約 1/3 がジェネラリスト，残り約 2/3 がスペシャリストに進む傾向がある．そのため，$K=2$ の場合には，CE の学生であっても文系的な志向を持つ学生は IS の学生と同一のクラスタに入ってしまう可能性があると考えられる．

$\lambda=0.0$, $K=2$ の場合の学生の分類結果を表 2 に示す．IS の学生は 1 名を除いて全て z_1 に分類されるのに対して，CE の学生は z_1, z_2 の両クラスに分類されていることが分かる．

表 2: $\lambda=0.0$, $K=2$ のときの学生の分類結果

	z_1（文系）	z_2（理系）	計
CE	48	77	125
IS	18	1	19
	66	78	144

分類誤り率=(48+1/144=34.0%)

各クラスの特徴文

表 3 に各クラスタの特徴文の例を示す．各クラスタの特徴文の傾向は概ね以下の通りである．

- 文系クラスタ（z_1）：
  主にコンピュータがどのように実社会でどのように役立っているのか，またコンピュータの利用方法などに興味がある．
- 理系クラスタ（z_2）：
  アーキテクチャやプログラミング，ソフトウェアなど応用技術の知識を学びたいと考えている．

(2) 隠れ属性数 $K=4$ の場合

$K=4, \lambda=0.0$ のときに誤分類率が最小となった．このときのクラスタリングの結果を表 4，各クラスタの特徴文を表 5 に示す．また図 2 はクラスタを併合していった際のデンドログラムである．

- z_1 が文系クラスタとなった．特徴文を見ると，コンピュータ・IT が実際企業でどのように使われているかということに興味がある意見が多い．文系志向の学生の特徴がよく出ているクラスタとなった．
- z_2, z_3, z_4 は理系クラスタである．コンピュータの中身やその応用技術に興味のある学生が多い．
- z_4 のクラスタではコンピュータのツールとしての側面に興味ある学生が多いように思える．

表 3: 隠れ属性 $z_k=2$ のときの特徴文

z	特徴文
z_1 (文系)	・[illegible]関心のあるコンピューター技術としては情報家電[illegible]のようなもの，従来それ単独での機能しかなかった家電製品がネットワークを介してほかの家電製品とつながることで新たな機能を生み出す発想[illegible]か，ネットワークの仕組みに興味があります ・コンピューターはどのように情報を取得することができ，また，知られるのを防ぐことが出来るのか ・IT が進化する現代社会では，コンピューターができなければ不利であると思う
z_2 (理系)	・[illegible]経営システム工学科[illegible]，コンピュータの使用方法や，実際にコンピュータを使ってさまざまなことをしましたが，この授業は，コンピュータを使用するというよりは，コンピュータについての知識を深めるための授業であるようなイメージがあります ・自分にとってはどちらかというと興味がある分野なのでこれを機にコンピュータの内部のことやあるいは C 言語や X プログラミング，そして，それにまつわることについて，さらには LINUX などの OS のことも少しは知識を増やしていきたいと思っています

表 4: $K=4$ のときの学生のクラスタリング結果

	z_1	z_2	z_3	z_4	計
CE	15	33	43	34	125
IS	16	2	0	1	19
	31	35	43	35	144

- このように同じ学科の中でもコンピュータに対する意識に違いが見える．このため，$K=4$ とすることにより，$K=2$ では分類できなかった学生の傾向を細分化することができた．分析のためには K を増やしたほうが，より深い視点で対象の違いを見ることができると考えられる．

表 5: $z_k=4, \lambda=0.0$ のときの特徴文

z	特徴文
z_1	・現在の情報化社会[illegible]企業行動を講義して[illegible]いるのを予想している ・実際に社会で，IT がどのように使われているのか，情報化社会において企業がどのように動いているのかが知りたくてこの講義をとりました ・企業にどのような影響を与えているのかについて，実際の現場からの話を聞かせてもらえる講義だと思い受講しました
z_2	・コンピューターに関することに対しては専門用語とかも全然わからないので，あまり専門用語を見たり聞いたりでのイメージですが情報通信技術，ネットワーク技術，ソフトウェア，ハードウェア，コンピューターグラフィックスなどが何か興味をそそられる感じです ・コンピューターのハードに関することだけでなく，普段あまりなじみのないハードウェアや，インターネット通信の仕組みやそれにかかわるセキュリティの技術などについても学んでいけると思う ・コンピューターが今どうやって科学技術の進歩に役立っているか[illegible]
z_3	・[illegible]このような授業があるように[illegible]理を学ぶことで終わってしまうかもしれないが，素人の私にとっては原理からしっかりと学び，その先を興味を持って取り組めるように講義で得た知識を自分のものにできたらいいと思います ・この科目は情報理論的な部分が多いので，自分にとってはかなり興味がある分野だと思いますが，教科書を見たところ難しそうだったのでついていけるか少し心配ですが，授業には全部出席して，きちんと聞いて，ちゃんと復習すれば理解できるような気がします
z_4	・そして，今では経営システム工学科に入学して，今までの遊び道具とは違い，パソコンを本格的な道具として捉えることになるための授業を受けている ・[illegible]で，大学で経営についての色々な勉強をし，企業経営全般についての幅広い経験と知識を手に入れ，深い洞察力と的確な判断力を培い，また分かりやすく説明する能力，優れた計算力も学び，大学在学中にインターンをして現場で働き自分を試し，将来，たくさんの企業から頼りにされる経営コンサルタントになりたいです

(3) 合成重み λ について

λ が 0.0 に近づく，すなわち自由記述式部分の重みが大きくなるにしたがって誤分類率が下がっていった．文書–索引語行列 A において，自由記述部分では索引語の出現頻度により重み付けされた値となっているが，選択式部分はすべての要素が $\{0,1\}$ をとっている．そのため，自由記述式と異なり特徴が出にくいのではないかと考えられる．あらかじめ重要な質問に高い重みを与えるなどの処理を行えば，選択式項目の特性をより反映させた有効なクラスタリングが可能になると考えられる．

5.3 日台の学生アンケート

インターネット等を活用することによって，国境の枠を超えて共通の講義を実施することも可能となっている．また，国際交流の活発化により，異なる国の学生同士が共通の講義を受講する機会も増えてきている．そのような状況で学生アンケートを実施する場合，選択式回答についてはあまり問題は生じないが，自由記述式については，学生は自国の言語を用いた方が回答

が負担にならず，また，細かいニュアンスまで正確に回答することが可能であるといえる．

そこで，本節では共通のアンケートに対して異なる言語で回答されたアンケートの分析手法の検討と結果の考察を行う．台湾では，立徳管理学院と淡江大学の情報系の講義において日本と同一内容のアンケートを実施した（回答数 107 名）．アンケート質問は人手により中国語に翻訳したものを使用した．また，分析の効率化のために中国語の回答データは中日翻訳ソフト (j 北京 v6 - http://www.kodensha.jp/soft/jb/) を用いて全て日本語に自動翻訳したものを人手で修正した．

前節で用いた日本の学生（144 名）と台湾の学生（107 名）の IQ をマージして文書-索引語行列 A を生成し，これまでと同様に PLSI によるクラスタリングを行う．

5.4 日台のアンケートに関する結果と考察

隠れ属性数が $K = 2, 3$ の場合のクラスタリング結果を表 6 に示す．

表 6: 日本・台湾学生のクラスタリング結果

$K = 2$

λ	0.0		0.5		1.0	
z_k	z_1	z_2	z_1	z_2	z_1	z_2
日本	0	144	0	144	118	26
台湾	90	3	102	5	24	83

$K = 3$

λ	0.0			0.5			1.0		
z_k	z_1	z_2	z_3	z_1	z_2	z_3	z_1	z_2	z_3
日本	0	83	61	0	86	58	15	68	61
台湾	85	4	4	90	4	13	79	19	9

(注)$\lambda = 0.0$ のとき台湾のデータの合計が 107 とならないのは，記述回答が全くの無回答（14 名）の学生があり，クラスタリングが不可能だったため．

(1) $K = 2$ の場合

・$\lambda = 0.0$：
ほぼ完全に日本と台湾でクラスが分割された．これは，自由記述文のみに着目してクラスタリングを行った際に，完全に日本と台湾クラスに分けられることを意味している．学生の傾向は国の違いによってもっとも顕著な差異があるということはごく自然なことであるといえる．ただし，クラスタリングにおいてスコアの高い単語（特徴語）を見てみると「方面」や「従事」など，日本の学生アンケートではあまり使用されない単語が高いスコアを持っていることがわかった．これは台湾の学生の特性というよりは日本語への自動翻訳の際に用いられる辞書に影響されていると考えられる．したがって，より自然な日本語へ翻訳されればこれらの単語の影響が排除され，学生の特性を探る上でより有効な分析結果が得られる可能性がある．

・$\lambda = 0.5$：
このときも概ね日台でクラスが完全に分割されており，選択項目を考慮した上でも $K = 2$ としたときには国の違いによってクラスタリングされることが分かった．

・$\lambda = 1.0$：
選択項目のみを見た場合には，日台ではきれいに分割することができない．すなわち，学生の特性が自由記述式回答のときほど日台の国の違いにあるわけではないことを示している．このとき，各クラスタの傾向は特徴語（選択項目もアルゴリズム上では語として扱われる）を見ることによって探ることが可能である．このときの各クラスタの特徴語を表 7 に示す．

(2) $K = 3$ の場合

表 7: 各クラスタの特徴文 ($K = 2, \lambda = 1.0$)

	特徴語
z_1 (台湾) など	UNIX について積極的に勉強したい=3（選択項目）， ネットワーク技術について積極的に勉強したい=3（選択項目）， 情報検索について積極的に勉強したい=3（選択項目）， 情報通信技術について積極的に勉強したい=3（選択項目）， 次の内容について積極的に勉強したい=3（選択項目）
z_2 (日本) z_2	この講義には毎回出席するつもりである=5（選択項目）， ウェブページ作成について積極的に勉強したい=5（選択項目）， EXCEL，WORD について積極的に勉強したい=5（選択項目）， ネットワーク技術について積極的に勉強したい=3（選択項目）， 興味ある科目については努力する=5（選択項目）， この講義を理解したいと思う=5（選択項目　など

表 8: 各クラスタの特徴語 ($K = 3, \lambda = 0.0$)

	特徴語
z_1 (台湾)	コンピュータ，方面，先生，概論，プログラム，設計， 課程，従事，
z_2 (日本 A)	パソコン，興味，講義，経営，分野，勉強，コンピュータ， 自分，システム，就職，インターネット，工学， 情報フィルタリングという用語を全く知らない（選択項目）など
z_3 (日本 B)	レポート，情報， ネットワーク技術を積極的に勉強したい=5（選択項目）， 情報通信技術を積極的に勉強したい=5（選択項=5 目）， 情報セキュリティ技術を積極的に勉強したい（選択項目）， ソフト・ハードウェアを積極的に勉強したい=5（選択項目）など

・$\lambda = 0.0, 0.5$ のとき，ほぼ完全に z_1 が台湾の学生，z_2, z_3 が日本の学生とクラスタリングされた．

・$\lambda = 0.5$ のときの各クラスタにおける特徴語を表 8 に示す．このことから，選択項目によって日本の学生も 2 つのクラスに分割され，z_3 の方が専門的な内容に強い意欲を持っていることが分かる．

・$\lambda = 1.0$ のときは，$K = 2$ のときと同様に，国の違いによるものとは別の学生の特性によってクラスタリングされていることが分かる．

6 まとめ

本稿では，選択項目と自由記述式を統合的に分析できる PLSI を用いたクラスタリング手法により，日本と台湾の学生アンケートの分析を行った．

その結果，アルゴリズムによって学生の特性に応じてクラス分けされること，さらに，そのクラスの特徴を把握するものとして重要語や重要文が有用であることを示した．

学生の特性と授業満足度・成績との関係，日台学生アンケート分析における自動翻訳機械の性能，未知語の取り扱い方法などを明らかにするのが今後の課題である．

参考文献

[1] T. Hofmann,"Probabilistic Latent Semantic Indexing," Proc. of SIGIR '99, ACM Press, pp.50-57, 1999.

[2] S. Hirasawa, and W. W. Chu,"Knowledge Acquisition from Documents with both Fixed and Free Formats," Proc. of 2003 IEEE Int. Conf. on System, Man, and Cybernetics, pp.4694-4699, U.S.A., Washington DC, Oct. 2003.

[3] 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（情報科学技術フォーラム）論文集，vol.2, pp.83-84, 江別，2003 年 9 月．

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman,"Indexing by Latent Semantic Analysis," J. of the Society for Information Science, 41, pp.391-407, 1990.

[5] 平澤茂一，コンピュータ工学，培風館，2001 年．