口頭発表B③

eラーニングシステムを経由して提出された理数系レポートに対する定量的解析

宇野 剛史
徳島大学大学院ソシオ・アーツ・アンド・サイエンス研究部
田中 彰
香川証券（徳島大学大学院総合科学教育部平成24年度修了生）

1．はじめに

近年WWWやPCの普及に伴い，ICTを活用した教育形態であるeラーニングシステムに注目が集まっており，学校，塾や企業での人材教育に広く用いられている．eラーニングシステムの導入は教材の提供や課題の管理における効率化を可能にし，学習者が提出したレポートやアンケートなどは電子データとして得られる．電子データは再利用や分析が容易なことから，本研究では蓄積されたデータに対する解析手法の提案およびその結果を紹介する．

分析対象として，徳島大学で開講されたeラーニングシステムを採用する講義の一つである総合科学部総合理数学科数理科学コース選択科目の「プログラミング演習Ⅱ」および「モデリング理論」を扱う．両講義では授業にて課題を要求した回が存在することから，提出されたレポートを分析した結果を紹介する．各講義において提出されるレポート課題は次の特徴を持つ．

- 「プログラミング演習Ⅱ」：C言語によるコード化を学ぶ講義であることから，プログラミング言語による表記が文章に含まれる
- 「モデリング理論」：実際の様々な現象を数理モデルによって解析する講義であることから，数式による表記が文章に含まれる

本研究では，レポートの特性に応じた分析手法を提案し，その結果を紹介する．

2．研究に用いた解析手法

レポートの分析手法として，本研究ではデータ分析手法の一つであるテキストマイニングを用いた．テキストマイニングはコンピュータを用いて膨大に貯蓄されたテキストの中から有益な情報を探し出す技術である．本手法において，膨大に蓄積された定型化されていない文章の集まりはテキストデータとして扱われ，自然言語処理により分かち書き，係り受けや形態素により解析することで，単語やフレーズに分割して得られる出現頻度や出現件数を集計する．得られたデータに対して，クラスター分析や回帰分析やを適用して分析を行う．

クラスター分析とは，対象間の距離を定義して距離の近さによって対象をいくつかの均質な物のクラスターに分類する方法の総称である．本研究では，レポートに含まれる単語群を成績や頻度により分類することで，レポート課題の詳細な分析のために用いる．クラスター分析手法はクラスター感の距離をどのように定義するかによって分類される．今回扱うデータにおける手法の優劣については次章で述べる．また，回帰分析は従属変数と連続尺度の独立変数の間に式を当てはめ，従属変数が説明変数によってどれくらい説明できるのかを定量的に分析する手法である．本研究では，成績と単語群の関連を見いだすために用いる．回帰分析は説明変数の数および式の線形性によって幾つかの手法に分類されるが，以下では線形回帰の単回帰分析を用いている．

本研究では，テキストマイニングを実装するためのツールとして，TinyTextMiner(TTM)を使用した．形態素解析にはMeCab，係り受け解析にはCaboChaを用いた．また，クラスター分析および回帰分析においては，表計算においてはMicrosoft Excel，統計解析においてはRを使用した．

3．分析結果および考察

提出されたレポートに対してテキストマイニングを適用することで，分解された形態素の各々に対して使用頻度および平均評価値（形態素を含んだレポートの評価点に対する使用頻度の重み付け平均値）が得られた．次にクラスター総数を10 に設定したクラスター分析を行った．得られた結果を図 1 に示す．クラスター間の距離の定義方法として広く使われている 7 つの方法を予備実験にて比較検討した結果，ウォード法以外でのクラスターの構成は1つか2つのクラスターに形態素が集中している一方，ウォード法に関してはそのような傾向は見られなかった．クラスターを用いたデータの分析においては少数のクラスターに形態素が集中していないことが好ましいことから，本研究ではウォード法が最も適していると判断できる．

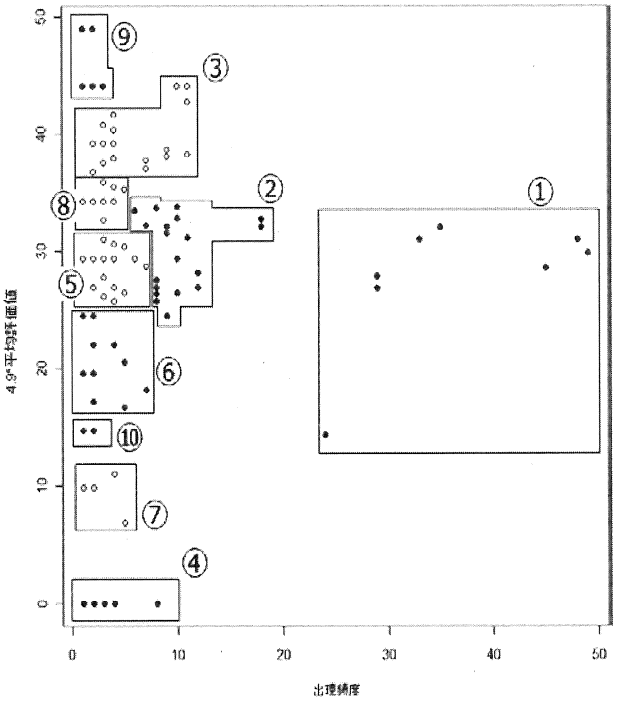

図 1　散布図上でのクラスターの構成（ウォード法）

まず，「プログラミング演習Ⅱ」のレポートについて述べる．上記で抽出した形態素は，全角文字，半角文字および全角文字と半角文字の混合の3 つの文字種に分類される．文字種において成績の平均値をとって比較したところ，半角文字の評価が他の文字種より低いことが分かった．図 1 におけるクラスター番号 4 をみると，評点が 0 点のレポート課題により抽出された語のみで構成されたクラスターが存在していた．語を確認すると半角文字によるプログラミング言語が多く含まれており，プログラムソースをそのまま提出していたことが評価が悪い要因であると考えられる．よって，このデータを異常値として除いた上で文字種別の点数の再度比較したところ，半角文字の評価が他の文字種の評価と比べて若干劣るがほぼ同程度の結果が得られた．以上より，本講義ではプログラミング言語での記述について指導が必要であることが分かる．

次に，「モデリング理論」のレポートについて述べる．「プログラミング演習Ⅱ」と同様に 3 つの文字種に分類できることから，前述と同様に文字種別に点数を比較したところ，半角文字の評価が他の文字種の評価より高いことが分かった．半角文字で構成される形態素を調べたところ，数式を表記したものが多く含まれると判明した．そこで，高い評点が得られた課題は数式を多く使用していると仮説を立て，文章中に含まれる数式を総バイト数と総数式数の 2 通りで回帰分析を用いて検証した．文章中に含まれる数式の記入方法として LaTeX を使用した数式の度合いは 1 とし，LaTeX を使用していない数式の度合いを $\alpha = 0.0, 0.2, \ldots, 1.0$ と与えた．その結果，全体的に総バイト数の相関係数が総数式数よりが高いことが分かった．出現総量の最高相関係数は重み $\alpha = 0.4$ の場合で，相関係数は 0.715 であった．以上より，特に LaTeX による数式表記の多いレポートは強い相関で成績が良いことが分かる．

4．おわりに

本研究では，徳島大学で開講された講義にて提出されたレポートに対して，テキストマイニングおよびクラスター分析を行った．さらに，専門用語の特殊性を考慮して文字種の差異に注目した回帰分析などによるより詳細な分析を行った．本研究で得られた知見は，様々な分野の e ラーニングシステムに対する定量的解析において広く応用可能であると考える．