

# アンケートデータに基づく情報リテラシーレベル分析

菅谷 克行

東京大学先端科学技術研究センター

sugaya@ai.rcast.u-tokyo.ac.jp

[概要] 近年、コンピュータ・通信技術の普及により情報活用能力の重要性は増している。その中で情報教育も新たな一歩を踏み出すべき時期に近付いてきている。しかしながら、情報教育の具体的な目標や内容等は未だ試行錯誤の域を脱しておらず、多角的視点からの分析や議論が必要とされている。そこで本研究では、情報リテラシーの履修学生を対象としたアンケート調査に基づき、各学生の情報リテラシーレベルを推定する手法及び分析した結果を報告する。特にアンケートデータを、特定の属性のみに注目した統計的手法のみではなく、多属性を同時に扱うことのできるデータマイニングの処理手法を用いて分析・考察を行った。

## 1. はじめに

近年、コンピュータ・通信技術の普及により情報活用能力の重要性は増している。更に新技術の開発は日進月歩であり、日々知識やスキルの更新が必要とされている。このような状況の中で情報リテラシー教育も新たな一歩を踏み出すべき時期に近付いてきている。しかしながら、情報教育の具体的な目標や内容等は未だ試行錯誤の域を脱しておらず、多角的視点からの分析や議論が必要とされている。

情報教育に限らず、教授内容や授業改善の検討・議論において、アンケートによる調査はこれまで多く実施されてきた。しかしながら、その分析は特定の属性のみに注目して集計・統計的処理を実施したものや、自由記述回答で意見をまとめたものなどが多く、アンケートデータの多属性を十分利用した分析は少ない。そのため、検討・議論が局所的で発展性のないものになってしまう場合も多かった。

そこで本研究では、情報教育に関する検討・議論が発展性を持つべく、新しい視点を提供することを目標に調査・分析を行った。コンピュータに馴染が薄いと考えられる文系学部女子学生に対する情報リテラシー教育に焦点をあて、アンケート調査と情報リテラシー能力を調査するためのテストを実施し、それらのデータ分析で得た結果から各学生の情報リテラシーレベルを推定する手法及び結果を報告する。特にアンケートデータを、特定の属性のみに注目した統計的手法のみではなく、多属性を同時に扱うことのできるデータマイニングの処理手法を用いて分析を行った。データマイニング手法を教育分野に適用する研究報告は近年増えつつあり、特に e-learning の学習履歴データの分析や運用に用いられている例が多い[1]。本研究では、結果の可読性に優れた決定木学習を利用して、情報リテラシー授業のアンケートデータから決定木を生成し、そこで得られたルールについて考察する。

## 2. 調査内容

### 2.1 調査対象およびデータ

本研究の調査対象は、女子大の文系学部で情報リテラシーの授業を履修している学生とした。文系学部生

対象の情報リテラシー授業では、学生のコンピュータ習熟度や利用歴、更には興味関心のばらつきが非常に大きく、教授内容や方法を工夫したり、学生のレベル分けをしたりすることは容易ではない。

また、分析に用いるデータは、授業で実施したアンケートデータとテスト成績である。アンケートデータは、履修者のこれまでのコンピュータ利用状況や知識に関する自己評価を、初回の授業時に回答してもらったものである。質問項目は、コンピュータの利用経験(ある・なし)、使用状況(情報授業・情報以外の授業・個人・その他)、自宅PC(ある・なし)、自宅PC利用(自分専用・主に自分が利用・家族で共同利用・利用しない)、自宅でのインターネット利用(可能・不可・分からない)、Web利用歴(Nヶ月)、Web利用頻度(週n回)、利用スキルの5段階自己評価(PC知識、キーボード、ワープロ、表計算、プログラミング、WWW、mail、html作成)、および初心者クラスを希望するか(はい・いいえ)である。

テスト成績とは、「Web検索課題の得点」、「定期試験の得点」、「タイピング試験の得点」の三種類である。「Web検索課題の得点」とは、Web上での情報の収集・選別の授業時に実施した2回の検索課題合計10問(20点満点)の得点である。「定期試験の得点」は、前期・後期の試験期間中に実施した、授業で扱った内容の理解度・定着度をみるための試験(80点満点)の結果である。そして「タイピング試験の得点」は、前期授業終了時に実施したキーボード入力テストの得点であり、10分間で指定された文章(日本語、英語)を正しく入力できた文字数である。本研究では、これら三種類の得点で情報リテラシーレベルを判断する。

### 2.2 分析手法

従来、アンケートデータの集計処理には、合計、平均値など統計的手法によってなされてきた。しかし、データベース技術の普及により大量のデータが蓄積されるようになると、それらのデータを有効活用する手法も研究開発されるようになった。その分野は、一般に「データマイニング(Data Mining)」、「データベースからの知識発見(KDD)」等と呼ばれる。手法の中心

は、主に人工知能の分野で開発された、事例データから知識（ルール、パターン）を発見する機械学習技術である。本研究で利用する決定木学習は機械学習技術の一つであり、データ中の注目する属性（目的属性）に関する知識を木構造によるルールの組み合わせで表現したものである。決定木学習の大きな特長の一つに可読性（Readability：生成した知識の読み易さ）がある。大量のデータから機械的に知識を発見するという事は、場合によっては意味のないルールや偶発的に抽出されてしまったルールが混じる可能性がある。そのため、発見された知識を、可読性によって人間が理解・確認できることは非常に大きな利点である。そのため、決定木学習アルゴリズムは多くの分析ツールに採用され、ビジネスの分野等で急速に広がった[2]。

本研究では、従来の統計的手法に加え、このような決定木による分析手法を利用して、学生の情報リテラシーレベルを分析・考察する。

### 3. 分析結果と考察

#### 3.1 テスト成績間の相関分析

まず三種類のテスト成績間の相関について分析した結果を述べる。事前の予想では、「情報リテラシーレベルの高い学生は、検索得点、定期試験得点、タイピング得点、総じて高得点であるのではないかと考えていた。しかし、分析の結果は、と、と、と、と、すべてにおいて相関係数は0.1以下であり、即ち相関関係は認められないという結論を得た。この結果は、学生の得意・不得意分野に統一性がなく、内容・レベル共に個人差が大きいことを示している。更に、テスト成績 ~ という三種類のデータから、情報リテラシーレベルを総じて判断することが困難であることも示唆している。たしかに授業の際、学生個人の理解内容に斑があり、個人差が非常に大きいと感じることも度々あった。本分析結果は、その状況を如実に示しているものと考えられる。この結果から、大人数を対象とした情報リテラシー授業の困難さが改めて浮き彫りになったと考えられる。

#### 3.2 決定木によるデータ分析

次に、三種類のテスト成績をそれぞれ目的属性とし、アンケートデータの各項目を条件属性として、決定木学習にかけて分析を行った結果を述べる。

本研究で用いた決定木学習のアルゴリズムは、様々なデータマイニングツールで一般的に利用されているC4.5[3]とし、生成された決定木は危険率0.25で枝刈りした。分析結果を目的属性毎に述べる。

##### 検索課題結果を目的属性とした場合

決定木によって抽出された条件属性は、「ワープロスキルに関する自己評価」、「自宅のPC利用環境」、「授業で初心者クラスを希望するかどうか」、「主なPC使用の状況」、「WWWの利用歴」であった。即ち、これらの属性の組み合わせ（フローチャート型）によって、検索得点の上位者と下位者の推定ができるということである。ここで抽出されたルールは、全体的に納得できるものである。例えば、「ワープロスキル」の自己評価が高いの方が、検索得点も高かったというルールは、

学生にとって、コンピュータスキルの中でも利用頻度が高いと考えられる「ワープロ操作」に自信がある、ということが検索スキルを推定する一つの鍵であることを示している。また、「自宅PCの利用環境」や「主なPCの使用状況」という属性ルールは、デジタルデバイスに代表される、「PCに触れることのできる環境」の重要性を示唆している。たしかに一般家庭にPCは浸透したとはいえ、学生個人で自由に利用でき、かつネットワーク環境も自由に使える者と、そうでない者の間に差が生じるということは、当然の結果であると考えられる。教育機関では、この点に関する積極的な配慮をするということも今後必要になってくるであろう。

##### 定期試験結果を目的属性とした場合

この分析からは、納得できるルールは得られなかった。例えば「プログラミングスキルの自己評価が2以下の者が、定期試験では上位」というように、偶然、パターンが一致してしまったと考えられるものが抽出されてしまった。定期試験の平均点が非常に高く、得点による人数のちらばりが少なかったことが、分析がうまく行かなかった原因だと思われる。

##### タイピング結果を目的属性とした場合

この分析では、「PC上でのメール操作に関する自己評価」が高いの方がタイピングも上位であるというルールを得ることができた。日常的にPCでメールの送受信を行っている者が、キーボード入力も速いということは十分納得できることであり、このルールがタイピング能力推定の鍵となり得ることが示された。しかしながら、この分析と同様、納得できないルールも抽出されている。この点は、無意味なルールの削除を可能とする、決定木学習手法の可読性の高さという長所を確認する良い機会となった。

### 4. まとめ

本稿では、統計的手法と決定木学習手法を利用して、情報リテラシー授業履修者のアンケートデータと成績データから、学生の情報リテラシーの分析と、その分析で得られたルールによって、情報リテラシーレベルを推定する手法について報告した。テスト成績間の相関分析では、情報リテラシーレベルを総じて判断することの難しさと、大人数での情報リテラシー授業の困難さが再認識できた。またデータマイニング手法によって得られたルールは、多少の誤抽出と思われるものも含まれていたが、全体的に納得できるものであり、データマイニング技術を教育分野に適用できる可能性を大きく示したものであった。

### 参考文献

- [1] 松居辰則, 岡本敏雄, “e-learningの統合評価機構におけるデータマイニング” 日本教育工学会第19回全国大会講演論文集, pp.583-586, 2003.
- [2] Witten, I.H., and Frank, E., *Data Mining*, Morgan Kaufmann, 2000.
- [3] Quinlan, J.R., *C4.5 Programs for Machine Learning*, Morgan Kaufmann, 1993.