シラバスデータのクラスタリングに基づく教育コース分析システムの構築

野澤 孝之† 井田 正明† 芳鐘 冬樹† 宮崎 和光† 喜多 一‡

大学評価・学位授与機構†
京都大学 学術情報メディアセンター‡

1．はじめに

各大学が独創的な教育コースを設計しようとする場合や，第三者が大学の教育コースの特徴を評価する場合，多数の大学にまたがる講義内容の横断的な把握が必要である．これは専門家にとっても負荷の高い課題であり，教育コース設計や評価の方針を立てるためのコンピュータ支援環境が望まれる．

本研究では，XML 化されたシラバスデータを対象に，それらが含む専門用語の出現頻度に基づき講義間の類似度を計算し，クラスタリングを行うシステムを構築した．大学一学科別，必修／選択の区分，履修年次などの軸に沿って講義のクラスタへの帰属分布を視覚化することで，教育コースを様々な角度から分析することを可能とした．

2．本システムによる分析の流れ

本システムが対象とするのは，井田らの提案する XML データ形式[1]に変換されたシバラスデータの集合である．分析は，1)含まれる専門用語にもとづき各シラバス（講義）の内容を縮約・定量化し，2)シラバス間の類似度を計算，3)類似度に基づきクラスタリング，そして 4)各シラバスのクラスタへの帰属分布を観察する，という手順で行われる．また，この各ステップで分析の詳細を規定する様々なオプションが出てくる．以下に分析の流れを説明する．また分析のオプションを表１に整理する．

2.1 シラバス（講義）内容の定量化

シラバス XML データについて，「科目名」，「授業概要」，「授業計画」などの項目の値が含む専門用語を抽出する．これにより，一つのシラバス s_i は

$$s_i = \{(term_k, score_{ik})\}_{term_k \in T_i} \quad (1)$$

という表現で定量化される．ただし T_i はシラバス s_i が含む専門用語の集合， $score_{ik}$ は s_i における専門用語 $term_k$ の重要度を表すスコアであり，利用する専門用語抽出手法により与えられる．

対象データ項目の文字列からの専門用語抽出手法としては，語の出現に関する統計を利用するものが一般的である．本システムでは，語の代表性と識別性に基づく手法（TF-IDF）[2]，連接頻度に基づく手法[3]を利用可能とした．

2.2 シラバス間の類似度計算

1.シラバス定量化	用語抽出の対象とするデータ項目 用語抽出手法(用語スコア計算含む)
2.類似度計算	規格化の有無 類似度の定義式
3.クラスタリング	クラスタリング手法
4.結果の観察	クラスタ帰属分布を比較する分類軸

表 1 分析のオプション

シラバス間の類似度を計算する前に，上ステップで得た各シラバスの定量表現を規格化するか否かを決定する必要がある．シラバスには記述量のばらつきがあり，記述量の多いシラバスほどより多くの専門用語と高いスコア合計を持つ傾向がある．そこで，全ての講義は同程度の内容を含むはずだと考える立場では，次の規格化操作

$$score_{ik} \leftarrow \frac{score_{ik}}{\sum_{term_l \in T_i} score_{il}} \quad (2)$$

を全てのスコアに対して施す．シラバスの記述量が講義内容の量を反映すると考える場合には，規格化は不要である．

規格化の有無を選択したうえで，全シラバス間の類似度を計算する．二つのシラバス s_i と s_j の類似度 $sim(s_i, s_j)$ の定義には様々なものが考えられるが，本システムでは専門用語スコアの重なりによる定義；

$$sim(s_i, s_j) = \sum_{term_k \in T_i \cap T_j} \min(score_{ik}, score_{jk}), \quad (3)$$

および全専門用語 $\bigcup_i T_i$ が張る空間におけるユークリッド距離の逆数による定義

$$sim(s_i, s_j) = \left\{ \sum_{term_k \in \bigcup_i T_i} (score_{ik} - score_{jk})^2 \right\}^{-1/2} \quad (4)$$

（ただし $term_k \notin T_i$ のとき $score_{ik} = 0$ と定義）を利用可能とした．

2.3 講義のクラスタリング

得られた類似度を用いてシラバスのクラスタリングを行う．ここでは，要素（シラバス）が座標空間内に埋め込まれていなくとも，要素間の類似度のみを用いてクラスタリングを行える手法が必要である．本システムでは，階層的クラスタリング手法[4]のうち最短距離法，群平均

Construction of Curriculum Analyzing System based on Clustering of Syllabus Data
† Takayuki Nozawa, Masaaki Ida, Fuyuki Yoshikane, Kazuteru Miyazaki
National Institution for Academic Degrees and University Evaluation
‡ Hajime Kita
Academic Center for Computing and Media Studies, Kyoto University

距離法，最長距離法を利用可能とした（効率化のため，ヒープを用いるアルゴリズム[5]を利用した）．

2.4 クラスタへの帰属分布への観察

以上の手続きで，シラバスは少数個のクラスタに分類される．各クラスタがどんな意味内容を持つかを把握する手掛かりとして，本システムは各クラスタの成立に強く寄与している専門用語のリストを提供する．専門用語 $term_k$ のクラスタ C_m の成立への寄与度 ctr_{km} は，次式で計算した；

$$ctr_{km} = \sum_{s_i \in C_m} score_{ik} \ . \qquad (5)$$

個々のシラバスがどんなクラスタに分類されるかだけでなく，大学一学科別，必修／選択の区分，履修年次などの分類軸（XML データ定義の中のカテゴリカルな項目）に沿ってシラバスのクラスタへの帰属分布を比較できると，教育コースの特徴をより把握し易くなる．本システムでは，上の三つの分類軸のうち一つを選択し，その分類軸を行，所属クラスタを列とするクロス表を作成する．このクロス表に主成分分析またはコレスポンデンス分析[6]を適用し，選択した分類軸の各ケースおよび各クラスタを平面上にマッピングできるようにした．

4．分析結果の例

情報工学系学科を対象に，Web を通じて収集した 16 大学 17 学科のシラバス（総数 1084，2002 年度版）を分析した結果の例を示す．なお分析のオプションは，1)「授業概要」「履修により達成される目標」「（授業計画）トピックス」を対象項目として TF-IDF に基づく手法で専門用語を抽出し（抽出された用語総数 17545），2)式(2)の規格化を施した上で式(3)の定義を用いてシラバス間の類似度を計算，3)群平均距離法で階層的クラスタリングを行い，4)大学－学科別の分類でクラスタ帰属分布を比較した．このオプション設定では特徴的なシラバスを抽出し易いが，クラスタのサイズもばらつきがちになる．

シラバス集合を７つのクラスタに分けたときの，各クラスタ（C1～C7）成立への寄与度の高い専門用語を表２に示す．クラスタ C1 は情報工学系の一般的なトピック，C2 はセミナーや演習，卒業研究など特殊な講義形態のもの，C3 はコンピュータグラフィックス（CG）関係,…等々のことが読み取れる．

また，大学一学科別でのシラバスのクラスタへの帰属分布を相対尺度的に規格化したうえで主成分分析を適用し，各大学－学科（A～Q）を平面上にマッピングした結果を図１に示す．これより，大学一学科 N や B は感性工学や CG の教育分野を含むこと，P は文化施設に関する独自の教科を持つこと，その他は類似していること等が読み取れる．

C1	回路，システム，関数，方程式，論理，計算，…
C2	研究，指導，技術，研究室，こと，プログラミ…
C3	ＣＧ，座標，変換，ベクトル，補間，レイトレ…
C4	経営，企業，ベンチャー企業，説明，経営者，…
C5	伝導，現象，超伝導，現象論，効果，理解，伝…
C6	感性，工学，感性表現，感性評価，情報，情報…
C7	文化会館，美術館，文化，博物館，文化政策,…

表 2　クラスタの形成に貢献した専門用語

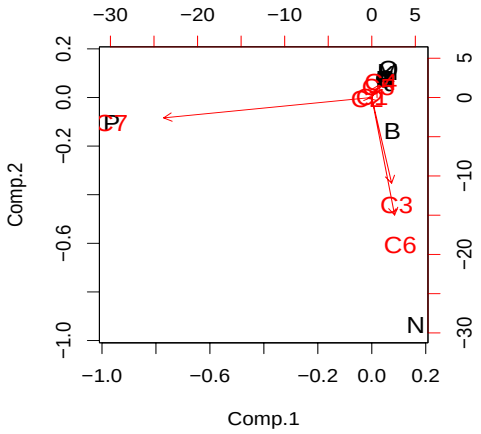

図 1　大学一学科別教育コースの特徴分布

5．考察

3 節で示したように，本システムの教育コース分析には多数のオプションがある．そしてオプションの選択によって分析結果の様相はときに大きく変わってくる．このような分析結果の不定性・多様な解釈の可能性は，複雑な対象からの知識発見にはつきものであり，むしろ教育コース設計者や評価者の視点を反映した幅広いオプションを試しては分析結果を観察するという繰り返しが，多角的な視点からの教育コース理解には必要であろう．

このような繰り返し分析を効率良く行ううえでは，対象データの操作や分析オプション設定のためのインターフェースの洗練，および分析オプション決定から結果取得までのターンアラウンドの短縮が重要である．そのため今後はデータベースとの連携，クラスタリングの高速化などを進めていく予定である．

謝辞　本研究を遂行するにあたりご協力いただいた大学評価・学位授与機構「大学評価情報の構造解析と評価プロセスへの応用の研究会」参加者の皆様に謝意を表します．

参考文献

[1]井田，宮崎，芳鐘，喜多：シラバスＸＭＬデータベースシステム構築に関する考察，情報処理学会第 65 回全国大会 2A-6，pp.4-247-4-248，2003.
[2]小西：自動構築型知識に基づく専門用語形成システム，情報処理学会論文誌，Vol.30，No.2，pp.179-189，1989.
[3]湯本，森，中川：出現頻度と連接頻度に基づく専門用語抽出，情報処理学会第 145 回自然言語処理研究会，pp.111-118，2001.
[4] 宮本定明：クラスター分析入門 ： ファジィクラスタリングの理論と応用，森北出版，1999.
[5] T. Kurita: An efficient agglomerative clustering algorithm using a heap, Pattern Recognition, Vol.24, No.3, pp.205-209, 1991.
[6]大隅ほか：記述的多変量解析法，日科技連出版社，1994.