

LDA と Isomap を用いた 計算機科学関連カリキュラムの分析

関谷 貴之^{1,a)} 松田 源立^{2,b)} 山口 和紀^{1,c)}

受付日 2012年2月13日, 採録日 2012年10月10日

概要: 大学の授業の体系性を維持する要となるのがカリキュラムであり, カリキュラムを改良するためには, 過去のカリキュラムや他大学のカリキュラムとの比較も必要となってくる. しかし, カリキュラムの比較は, 教員がシラバスなどの内容を読んで比較しなければならず簡単ではなかった. 本研究では, 確率的な文書モデルである LDA と次元圧縮手法である Isomap を用いて, 授業の内容を説明するシラバスからマップを生成し, カリキュラムを比較する手法を提案する. この手法を用いて 2つの学科の情報系のカリキュラムを分析することで, 本手法の有効性を評価する.

キーワード: カリキュラム, シラバス, LDA, Isomap

Analysis of Computer Science Related Curriculum on LDA and Isomap

TAKAYUKI SEKIYA^{1,a)} YOSHITATSU MATSUDA^{2,b)} KAZUNORI YAMAGUCHI^{1,c)}

Received: February 13, 2012, Accepted: October 10, 2012

Abstract: A curriculum is crucial to maintaining the integrity of various courses of universities, and an effort to improve a curriculum may include an analysis of the curriculum by comparing the curriculum with previous curricula or curricula of other universities. However, comparing them by hand is too labor-intensive. In this paper, we propose a method to project syllabi on a map by latent Dirichlet allocation (LDA) and Isomap for the comparison of curricula embodied by the syllabi. We applied this method to information science-related curricula of two departments and evaluated the effectiveness of the method.

Keywords: curriculum, syllabus, LDA, Isomap

1. はじめに

大学の授業の体系性を維持する要となるのがカリキュラムである. カリキュラムに関する研究や取り組みは, 国家単位でのカリキュラムの開発や, 教育現場でカリキュラムを作成する際の原則の提唱, カリキュラムの改良方法など, 様々なものがある [15]. 本論文では, カリキュラムの理念

ではなく, BOK (Body of Knowledge) や授業などで具体化されたカリキュラムを対象とし, それらを把握しやすく可視化することで, カリキュラムの改善につなげることを目的とする.

具体化されたカリキュラムとしては, カリキュラム標準や学部における授業群のようなものを考える. たとえば, 計算機科学分野におけるカリキュラム標準としては, ACM や IEEE などの学会が提案している Computing Curricula [1], [21] や情報処理学会が提案しているコンピュータ科学カリキュラム標準 J07-CS [24] などがある. これらのカリキュラム標準では, 「アルゴリズム」や「オペレーティングシステム」などの分野全体を内容に応じて分類した「領域」が定義されている. 授業自体は文書ではないの

¹ 東京大学
The University of Tokyo, Meguro, Tokyo 153-8902, Japan

² 青山学院大学
Aoyama Gakuin University, Sagami-hara, Kanagawa 252-5258, Japan

a) sekiya@ecc.u-tokyo.ac.jp

b) matsuda@it.aoyama.ac.jp

c) yamaguch@graco.c.u-tokyo.ac.jp

で、本論文では、授業のシラバスを、授業を表す文書と考えて分析の対象とすることにした。授業に対する分析の材料としてシラバスを用いることは、ウェブ経由で集めたシラバスのリポジトリを作成して、他の大学や教育機関で行われている授業を分析したり [22]、自組織のシラバスを分析したうえでそれを分かりやすく表示したりする [12] などでも行われている。

筆者らはカリキュラムを把握しやすくするために、授業を表すシラバスを相互の関係に従って配置したマップを生成する方法を研究している [18], [19]。マップを可視化することにしたのは、カリキュラムに知識のある人が読み取ることで、その人の知識を活用し、機械的には抽出が困難な情報も読み取ることができると期待してのことである。

マップ生成のための手法としては、確率的な文書モデルである Latent Dirichlet Allocation (LDA) [5] と次元圧縮手法である Isomap [20] を用いている。LDA を用いることで、カリキュラム標準に定義された領域をトピックとして抽出し、抽出したトピックを用いて、一般のカリキュラムのシラバスをカリキュラム標準の領域に基づくトピック空間の中に配置することができる。また、Isomap を用いることで、空間中の局所的な関係をできるだけ保存したまま、全体を平面上に射影することができる。

実験では、J07-CS で抽出したトピック空間に基づいて、2つの学科の情報系のカリキュラムを J07-CS 上で比較し、カリキュラムの特徴を読み取った。次に、他のシラバスと所定の関係になるようシラバスを書き換えるためにマップを活用する実験を行った。

以下、2章では関連研究をあげて本研究と比較する。3章ではカリキュラム分析手法の詳細について述べ、4章では本研究の手法を用いたカリキュラム分析の結果を報告する。5章はまとめである。

2. 関連研究

カリキュラムに関する研究としては、カリキュラム設計ツールが開発されてきており、そのいくつかは公開されている [8], [23]。そのようなツールのほとんどでは、カリキュラムを設計する教員がユニット化された知識の断片 [14] を定義する必要があり、ツールを利用するのに非常に手間がかかる。

Tungare らは計算機科学に関するシラバスのリポジトリを構築し、シラバスどうしを比較する SyllabusMaker と呼ぶシステムを開発している [22]。しかし、本研究で行っているようなカリキュラム全体を把握する仕組みは用意されていない。

教職員や学生にとって、授業に関する基本情報はシラバスで提供される。そこで、シラバスを分析する研究もいくつか見られる。たとえば、Ronchetti らは CC1991 と Wikipedia を用いたシラバスの比較分析を計画している [16]

が、結果は示されていない。

Mima らは、自動的に専門用語を認識してクラスタリングする技術を用いた MIMA search [12] やシラバス検索のための分類システム [13] を開発している。さらに、東京大学の講義資料を公開する UT OpenCourseWare^{*1}では、MIMA search を用いた「シラバスの見える化」を行い、専門用語に基づいたシラバスどうしの関係を示している。井田らは対話的にカリキュラムを分析するツールを開発している [9], [10]。しかし、いずれのツールにも、本研究で行っているようなカリキュラムの構造の可視化の機能は備わっていない。

Carvalho らは、欧州や北アフリカ諸国の小中高校のシラバスを比較分析している [6]。この研究では、教員や生徒に対するアンケートに基づく分析を実施するために、予備的なアンケート結果の回答を MCA (Multiple Correspondence Analysis) で分析することで、その特徴を読み取れることを試みている。対象をシラバスとする点や分析結果を可視化する点で、本研究との関連がある。しかし、分析の材料がアンケートの回答であることから、回答者の主観による結果のばらつきや偏りが生じるおそれがあり、これを避けるためには、多くの回答を集めて分析する必要がある。本研究は、学会などが定めたカリキュラム標準を基準とした統計的な処理結果に基づいて、カリキュラムを分析することで、主観的な要因が含まれることを排し、かつ比較的簡便に分析結果が得られる。

カリキュラムに限定せず、大量の文書からなる文書群全体を可視化する方法としては、生成モデルに基づいて文書から直接 2次元平面にマップするもの [11] がある。しかし、異なる文書群を同一マップ内で比較するのは困難である。それに対して本研究では同一のマップに複数の文書群を射影することで、文書群の比較を行うことができる。

3. 分析手法

3.1 分析の要件

本論文では、多数の BOK やシラバスで表現されたカリキュラムをシラバスの集合ととらえ、そのシラバスを平面上に記号として配置したものをカリキュラムのマップと呼ぶ。そして、カリキュラムの特徴は、このマップ上の記号の分布として視覚的に把握できるものと仮定する。さらに、異なる大学のカリキュラムや過去のカリキュラムとの比較を行うことを考えると、複数のカリキュラムを1つのマップの上に乗せられることが望ましい。そこで、本研究では、以下の3つの条件を満たす可視化を提案する。

- (A) 与えられたカリキュラムを基準として可視化が行えること
- (B) シラバスの関連度がマップ上の距離として反映されること

*1 <http://ocw.u-tokyo.ac.jp/>

(C) 異なるカリキュラムを同一の基準でマップできること

3.2 LDA

3.1 節で述べた条件 (A) を満たすために、与えられたカリキュラム標準の個々の専門的な領域に合致したトピックを LDA を用いて抽出し、そのトピックをカリキュラム分析の基準として用いることにした。LDA とは、確率的文書モデルに基づいて文書の集合から個々の文書の特徴づける潜在的なトピックを抽出する手法である。カリキュラム標準においては、「領域」としてトピックがあらかじめ定義されているが、本研究では、それらのトピックを LDA の確率的モデルとして再抽出する。これにより、カリキュラム標準に含まれない任意のシラバスが、カリキュラム標準の各トピックにどの程度関連するかを計算できるので、各シラバスを共通のトピック空間上に配置することが可能となる。LDA の確率モデルの概要を以下に説明する。詳細は Blei らの論文 [5] を参照してほしい。

- (1) 前提として、LDA では文書を確率的に選択された単語の集合 (bag of words) ととらえる。そこで、文書 $w = (w_m)$ ($m = 1, \dots, M$) を、 N 種類の語彙 v_1, \dots, v_N が M 個の単語 w_1, \dots, w_M として現れる多重集合 (同一文書中に単語として同一の語彙が複数回登場することを許す集合) として扱う。すなわち、 $w_m \in \{v_1, \dots, v_N\}$ である。
- (2) 最初に、Dirichlet 分布 $\text{Dirichlet}(\alpha)$ に従って、文書の特徴づけるトピックの確率ベクトル $\theta = (\theta_t)$ ($t = 1, \dots, T$) が決定されると仮定する。ここで、トピックは T 種類、 α は T 次元 Dirichlet パラメータである。 $\sum \theta_t = 1$ である。
- (3) 次に、文書 w に含まれる m 番目の単語 w_m のトピック z_m ($m = 1, \dots, M$) が、 θ の多項分布に基づいて決定される。 $z_m \in \{1, \dots, T\}$ である。
- (4) 最後に単語 w_m が、条件付き多項分布 $p(w_m | z_m, \beta)$ に従って語彙の中から選択される。ここで、 $\beta = (\beta_{nt})$ は $\sum_n \beta_{nt} = 1$ を満たす多項分布パラメータであり、トピック t と語彙 v_n の関連の強さを示す。

テキストコーパスにおいて、コーパスを構成する個々の文書に含まれる単語の頻度情報を LDA で処理すると、コーパス中の文書はそのトピックの確率ベクトル θ に基づいてトピック空間内で特徴づけられる。つまり、カリキュラムをシラバスの集合と見なして、LDA で処理することで、個々のシラバスを θ を用いて T 次元上で特徴づけることができる。

LDA においては、文書 w の集合のみが既知であり、他のパラメータは何らかの基準で決めるか、推定して決める必要がある。推定および設定手法として、変分ベイズ法 [5] や Gibbs サンプリング法 [7] などが提案されているが、本

研究では LDA の提案者である Blei らによって構築され、広く使われている変分ベイズ法を用いた。

LDA におけるパラメータは、大きく分けて、生成モデルを規定する基準となるパラメータ (α, β) と、ある与えられた文書がどのトピックに関係しているかを示すパラメータ (θ, z_m) の 2 種類がある。基準となる文書集合をもとに (α, β) を推定し、それ以外の文書に関しては (α, β) を固定して (θ, z_m) を推定することで、与えられたすべての文書を基準パラメータによって特徴づけることができる。ただし、 (θ, z_m) の推定は膨大な積分計算を必要とするため、一般に困難である。変分ベイズ法では、この推定を近似的に解決している。本論文では、 θ に関してのみ説明する。

θ については、 θ の分布 $P(\theta)$ の推定問題を、 $P(\theta) \simeq \text{Dirichlet}(\theta | \gamma)$ で近似し、近似の最も良くなる T 次元 Dirichlet パラメータ $\gamma = (\gamma_t)$ を推定する問題に帰着させる。 γ は Dirichlet パラメータであり、 θ そのものの値ではないが、Dirichlet 分布の性質により、 θ の期待値 $E(\theta)$ は、合計を正規化した $\bar{\gamma} = \gamma / \sum \gamma_t$ で与えられる。そこで、本研究では、この $\bar{\gamma}$ を、文書のトピック空間での位置と解釈する。ベクトルである $\bar{\gamma}$ のトピック t に対応する座標値が、その文書とトピック t との関連の強さを表す。

本研究では、 $\alpha = (\alpha_t)$ のすべての要素の値は一定値 α であると仮定する。これは、LDA で抽出されるトピックの生成率、つまりトピックに対応するカリキュラム標準の個々の専門的な領域の存在に偏りがないと仮定することに対応する。J07-CS における与えられた各トピックは代表的なものが偏りなく提案されているので、この仮定は妥当である。

3.3 分析方法

本節では、LDA と Isomap を用いたカリキュラムの分析手法を説明する。分析には、次の 3 種類のカリキュラムを用いる。ここでは、カリキュラムはシラバス w_i の集合で表されている。

Target Curriculum, $C_{\text{target}} = \{w_i^{\text{target}}\}$: 本手法を用いて分析するカリキュラム。

Model Curriculum, $C_{\text{model}} = \{w_i^{\text{model}}\}$: C_{target} が対象とする学問分野を広くカバーする標準的、模範的なカリキュラム。その学問分野のトピックの抽出に用いる。

Reference Curriculum, $C_{\text{ref}} = \{w_i^{\text{ref}}\}$: C_{target} の分析にあたって比較基準となるカリキュラム。

図 1 に示す分析手法の概念図に沿って、分析の手順を説明する。

Step 1: Model Curriculum からのトピックの抽出
標準カリキュラム C_{model} を LDA で分析する (Topic Estimation) ことで、分析対象の学問分野におけるトピックの生成モデルのパラメータ β_{model} を推定する (本論文では α は既知としている)。以降の Step では、この β_{model} を固定して、分析対象となるカリキュラムの標準的トピック

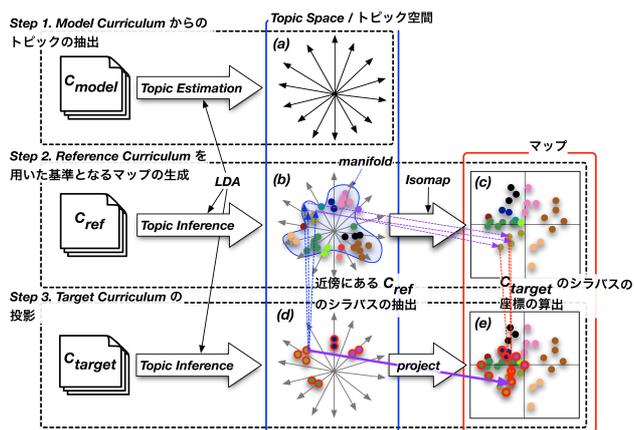


図 1 カリキュラム分析手法の概念図

Fig. 1 Conceptual diagram of curriculum analysis method.

空間として用いる (図 1(a)). 任意のシラバス w_i に同一のトピック生成モデル β_{model} を用いることができるので, 3.1 節の条件 (A) を満たす.

Step 2: Reference Curriculum を用いた基準となるマップの生成

まず β_{model} を利用して, C_{ref} の各シラバスの $E(\theta) = \bar{\gamma}(w_i^{ref})$ を LDA で推定する (Topic Inference). これにより, C_{ref} の全シラバスを, 基準となるトピック空間上に配置する (図 1(b)). しかし, トピック空間は T 次元空間であり, このままでは, カリキュラムの全体構造を可視化できない. そこで, 本論文では, 高次元空間内の多様体構造を保持したまま, これを広げるようにして次元を縮退する手法 Isomap [20] を用いる.

$$x(w_i^{ref}) = \Pi_{Isomap}(\bar{\gamma}(w_i^{ref})) \quad (1)$$

ここで, $\Pi_{Isomap} : R^T \rightarrow R^2$ は Isomap による写像変換である. また, $x(w_i^{ref})$ は各シラバスの 2次元平面上での位置を示す. この結果, 3.1 節の条件 (B) を満たすような C_{ref} を 2次元平面上に展開したマップが得られる (図 1(c)).

Step 3: Target Curriculum のマップへの射影

まず Step 2 と同様に β_{model} に基づいて, C_{target} の各シラバスの $\bar{\gamma}(w_i^{target})$ を LDA で推定する (Topic Inference). これにより基準トピック空間に C_{target} が配置される (図 1(d)).

次に, 各シラバス w_i^{target} を, Step 2 で求めたマップに, 以下の式を用いて射影する.

$$x(w_i^{target}) = \sum_{k=1}^{k_{neigh}} \eta_k x(w_k^{ref-neigh}). \quad (2)$$

ここで, $w_k^{ref-neigh} \in C_{ref}$ は, トピック空間上で $\bar{\gamma}(w_i^{target})$ の近傍となる k_{neigh} 個の比較基準シラバスである. 本論文では, $k_{neigh} = 3$ とした. η_k は $\sum \eta_k = 1$ を満たす加算重みであり, この制約下で $\sum \eta_k \bar{\gamma}(w_k^{ref-neigh})$ と $\bar{\gamma}(w_i^{target})$ の距離を最小化することで計算する. この計算は, 多様体

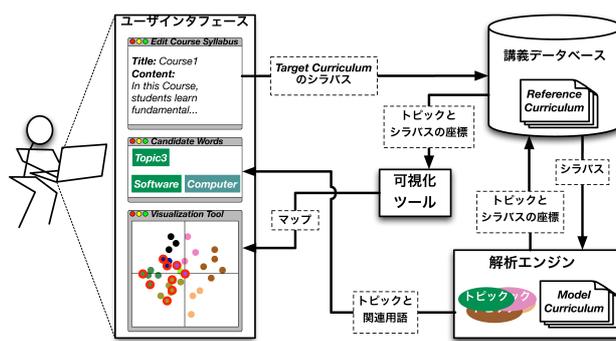


図 2 カリキュラム分析ツールの構成

Fig. 2 System architecture of curriculum analysis tool.

次元縮退法の一種である LLE [17] で行われるものと同一である. この結果, C_{target} のすべてのシラバスが, Step 2 で求めた C_{ref} のマップと同一の 2次元平面上に射影される (図 1(e)). 異なるカリキュラムを同一のマップ上で比較できるので, 3.1 節の条件 (C) を満たす.

4. 実験

4.1 システム構成

我々は, 3.3 節で述べた手法に基づいてカリキュラムを分析するツール群と, Step 3 における Target Curriculum の入力を支援する Web ベースのインタフェースを開発した. ツール全体は Perl CGI と PostgreSQL データベースで実装しているが, 行列演算処理では GNU Octave を呼び出すなど, 一部で Perl 以外の外部プログラムを利用している. 図 2 にツールのアーキテクチャを示す.

講義データベースは, 利用者がユーザインタフェースから入力したタイトルや内容などの Target Curriculum の情報を保存する. また, Reference Curriculum のシラバスなどの可視化に必要なデータ, つまり Step 1 と 2 を実行して得られたデータを保存する.

解析エンジンは 3.3 節の手法を実装したカリキュラム分析用のエンジンである. LDA の処理には, Blei のモデルを C 言語で実装した LDA-C [4] を用いる. 解析エンジンは, LDA を用いたトピックの抽出や各シラバスの座標値の算出を行うほかに, 利用者が入力したシラバスと指定したトピックに基づいて, 関連のある用語を提示することができる.

可視化ツールは解析エンジンで生成したマップを表示するツールである. Ajax の技術を用いて, 表示されたマップを拡大・縮小・移動することができる.

なお, Target Curriculum の 1 個のシラバスのテキストから, マップの表示に要するデータを取得までの所要時間は数秒程度である. この時間には, MeCab でシラバスのテキストから単語を抽出する時間や, LDA-C や Octave などの外部プログラムの実行時間が含まれる. また, Step 1 と 2 に対応するツールの実行に要する時間は, 本論文で扱った 140 個程度のシラバスからなるカリキュラムの場合は 4,

5分程度である。

本システムは主に次の2つの作業を支援する機能がある。(B)について4.6節で説明する。

A. カリキュラムの分析 利用者：分析したいカリキュラムを構成するシラバス文書を入力する。加えて、ツール上で事前に登録されたカリキュラムから、分析の基準となる Model Curriculum と Reference Curriculum を選択する。システム：選択された Model Curriculum から抽出された複数のトピックと当該シラバスとの関係の強さ、および当該シラバスを Reference Curriculum に基づくマップ上に投影した結果を示す。

B. シラバスの改良 利用者：カリキュラムマップ上の任意の点をクリックする。システム：Model Curriculum のシラバス文書に含まれる単語で、かつその点の近くにある Reference Curriculum のシラバスに含まれる単語を提示する。利用者：提示された単語を含むシラバス文書を入力して、指定した点の近くに配置されるシラバスを作成できる。

4.2 最適なパラメータの設定

一般に LDA においては、 α の値を 1 より小さくすれば、シラバスは 1 個あるいはごく少数のトピックとの関係が強くなり、1 より大きくすれば多数のトピックとの関係が同程度に強くなる傾向がある。その結果得られるカリキュラムのマップについては、 α が小さいとトピックに対応すると考えられるマップ上の狭い領域にシラバスが固まって配置され、後者では中央付近に固まって配置される。いずれにしても、シラバスどうしの違いや関係を視覚的には把握しにくくなる。このような観点から $\alpha = 1$ とすることが適切であると考えられる。

また、3.2 節で述べたように、本研究ではカリキュラム標準における領域を LDA のトピックとして抽出して、これを基準としてトピック空間を形成し、カリキュラムマップを生成することから、トピックの数 T は領域の数と同じ値にすることが適切であると考えられる。

このようなパラメータの設定が妥当か否かを検証するために、Dirichlet パラメータ α とトピック数 T を決める予備実験を行った。本研究で提案するカリキュラムの分析方法においては、LDA で抽出したトピックと文書との関連の強さが、Isomap を用いて得られるマップ上の文書の座標として適切に表現され、互いに関係のある文書がマップ上で近くに配置されるかが重要である。そこでこの実験では、あらかじめ分類された文書集合として UCI Machine Learning Repository^{*2} [2] が提供する Twenty Newsgroups Data Set (文書数：2,000 個、分類数：20 個) を用いた。文書群によって確率モデルが異なるため、本来はカリキュラ

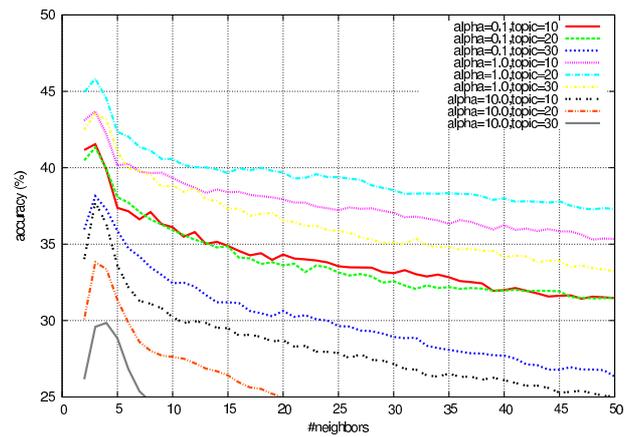


図 3 異なる k の値に基づく k 近傍の一致率 (パラメータ $\alpha = 0.1, 1, 10$, トピック数 $T = 10, 20, 30$)

Fig. 3 Accuracy of the k -nearest neighbor method for the different number of neighbors k for Dirichlet parameter $\alpha = 0.1, 1, 10$ and the number of topics $T = 10, 20, 30$.

ム標準自体をデータセットとして用いるべきである。しかしながら、今回カリキュラム標準として与えられている文書集合のサイズ約 140 個程度では、パラメータの最適値を求めるには十分ではないと考えられる。そこで、広く用いられている UCI Machine Learning Repository を用いて、あくまで目安ではあるが、安定的に良い結果を返すと期待される α と T の値を求めた。

3.3 節で述べた LDA と Isomap を使った可視化手法を、パラメータの値を変更しながら適用して、同じ分類の文書 (ネットニュースの記事) がより近くに配置されれば、そのときの値が適切だと判断する。判断指標として、Leave-one-out cross validation^{*3}に基づく k 近傍法による分類精度を用いた。すべてのパラメータ値について、同じ分析処理を 10 回繰り返して、得られた分類精度の平均値で評価した。 $\alpha = \{0.1, 1.0, 10\}$, $T = \{10, 20, 30\}$ のときの精度のグラフを図 3 に示す。この予備実験の結果から、トピック数 T の値を元データの分類数と同じ 20 かつ Dirichlet パラメータ α を 1 としたときの精度が高いことが分かった。そこで、当初想定した $T =$ カリキュラム標準の領域数および $\alpha = 1$ を採用することとした。

なお、3.3 節で説明した分析方法 Step 2 で用いる Isomap では、トピック空間内での近傍数によって平面上のマップが変化する可能性がある。しかし、本論文で対象としたカリキュラムを分析した限りでは、値の違いによってマップ上のシラバスの分布が大きく変わることがなかったことから、Reference Curriculum のすべてのシラバスを展開可能で最も小さい値を用いた。

4.3 Reference Curriculum の設定

ある学科の情報系のカリキュラムを分析するために、情

*2 入手先 (<http://archive.ics.uci.edu/ml/>)

*3 詳しくは、機械学習の教科書 [3]などを参考してほしい。

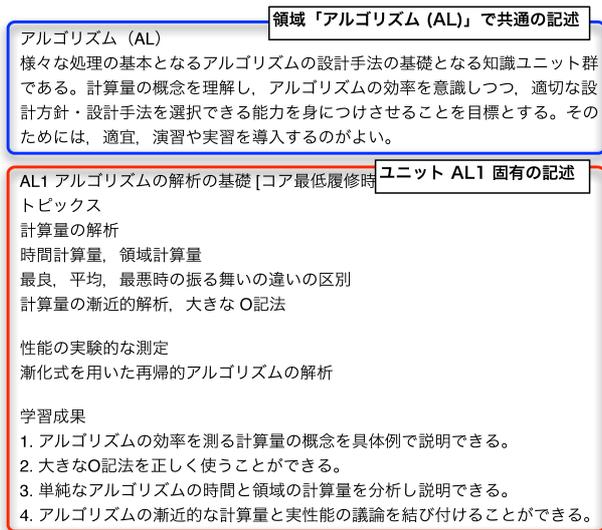


図 4 カリキュラム標準 J07-CS のユニットの記述に対する前処理
 Fig. 4 Pre-processing of the description of each “Unit” in J07-CS for Model Curriculum.

報処理学会が作成したコンピュータ科学カリキュラム標準 J07-CS [24] の BOK (Body of Knowledge) 全体を Model Curriculum とする。カリキュラム標準を Model Curriculum として利用するに際して、BOK = カリキュラム、領域 = トピック、ユニット = シラバスと解釈した。J07-CS の領域は 15 個、ユニットの総数は 138 個で、個々の領域のユニットは 4~14 である。ユニットの説明は比較的狭い範囲の事柄を列挙したもので、同じ領域に属する互いに関係のあるユニットであっても、共通の単語がないあるいは非常に少なく、LDA で抽出したトピックとの関連の強さがまったく異なる場合もありうる。3.2 節で述べたように、カリキュラム標準の各領域に合致したトピックを抽出するために、ユニットの説明にそのユニットが属する領域の記述を加えて分析し、領域を反映したトピックが抽出されやすいようにした。たとえば、領域「アルゴリズム (AL)」のユニット「AL1 アルゴリズムの解析の基礎」を処理する際には、領域共通の記述 (図 4 の上部の青い四角に囲まれた部分) とユニット固有の記述 (図 4 の下部の赤い四角に囲まれた部分) を合わせたものを、シラバス相当の文書とした。その結果、同じ領域に属するユニットは、LDA の同じトピックとの関連が強くなり、LDA のトピックと J07-CS の領域とをほぼ一致させることができた。

図 1 に示すように、本分析手法において Reference Curriculum はマップ全体の形状や配置のもととなることから、分析対象の範囲を広くカバーするカリキュラムが必要である。ここでは、Model Curriculum にも用いた J07-CS の BOK と、情報処理学会が一般情報教育用の教科書として作成した「情報とコンピューティング」*4 と「情報と社

*4 川合慧監修, 河村一樹編著, オーム社, 2004 年.

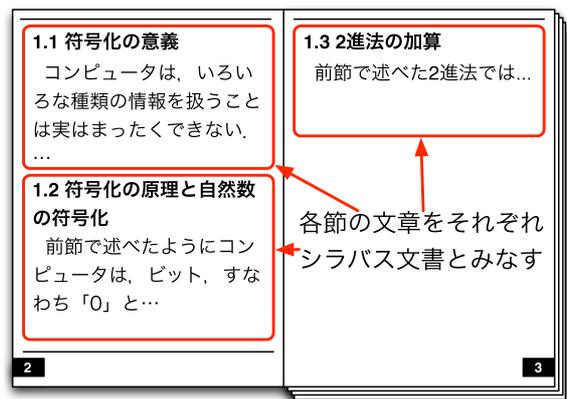


図 5 Reference Curriculum 用シラバス文書の抽出方法の概念図
 Fig. 5 A conceptual diagram of extracting syllabi from a textbook for Reference Curriculum.

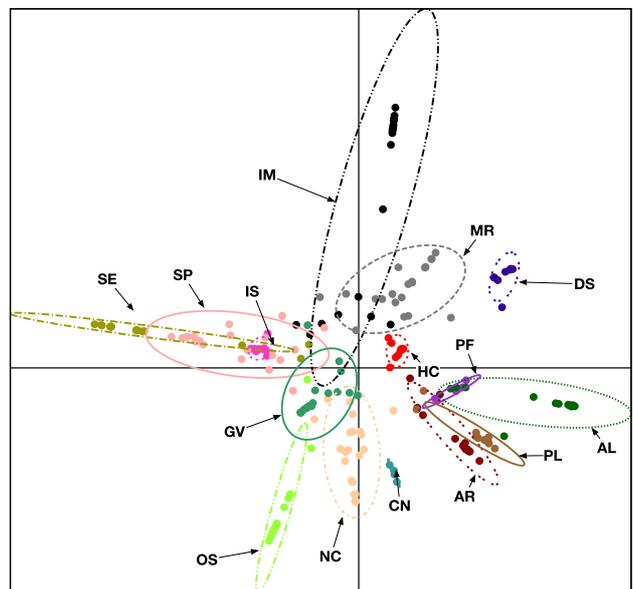


図 6 J07-CS (Reference Curriculum) のマップ
 Fig. 6 Map of J07-CS (Reference Curriculum).

会」*5を用いた。前者の J07-CS BOK については、Model Curriculum と同様にユニットをシラバス相当の文書とする。ただし、Model Curriculum の場合とは異なり、領域の記述をユニットに加える前処理は行わない。後者については、教科書全体からテキストを抽出して、図 5 に示すように、「1.1 符号化の意義」「1.2 符号化の原理と自然数の符号化」のような節の単位で分け、全部で 69 個の節をそれぞれ独立したシラバス文書として扱った。その結果 Reference Curriculum のシラバスの数は合計で 207 個となる。ここで得られた Reference Curriculum のマップを図 6 に示す。

1 つのシラバスは様々な内容を含むが、3.2 節で述べた LDA を用いるとシラバスとトピックとの関係は確率として求められるので、各シラバスに対して最も確率が高いトピックが決まる。このトピックを主たるトピックと呼ぶ。

*5 川合慧監修, 駒谷昇一編著, オーム社, 2004 年.

表 1 J07-CS のエリアとラベルとの関係

Table 1 Relationships between labels and areas of J07-CS.

ラベル	J07-CS の領域	マップ上の色
IS	インテリジェントシステム	ピンク
PF	プログラミングの基礎	紫
AL	アルゴリズム	緑
IM	情報管理	黒
PL	プログラミング言語	茶
HC	ヒューマンコンピュータインタラクション	赤
SP	社会的視点と情報倫理	桃
OS	オペレーティングシステム	黄緑
GV	グラフィックスとビジュアル・コンピューティング	青緑
NC	ネットワークコンピューティング	オレンジ
AR	アーキテクチャと構成	赤黒
SE	ソフトウェア工学	黄
CN	計算科学と数値計算	青
DS	離散構造	紫
MR	マルチメディア表現	灰

主たるトピックは、そのシラバスの主な内容を示すと考えられる。そこで、表 1 に示すようにトピックの色を決めて、図 6~図 9 のカリキュラムマップ上では、それぞれのシラバスの主たるトピックに応じた色の丸記号でシラバスを示す。これによって、Reference Curriculum のマップ上で同じ色の丸記号が固まって配置される部分は、表 1 に示す J07-CS の領域に対応すると考えられる。図 6~図 8 では、この分布を把握しやすくするために、マップ上のシラバスの丸記号の座標値の分散共分散行列から 2 つの軸を求め、 2σ ($\sigma^2 = \text{分散}$)*6 の楕円で囲んでいる。楕円を指し示すラベルは表 1 の J07-CS の領域に対応する。

分析対象である Target Curriculum のシラバスは、Reference Curriculum のマップ上に反映されて、主たるトピックに対応する楕円の中やその近傍に配置されることが期待されるので、Reference Curriculum のマップが適切に生成されている必要がある。そこで、Reference Curriculum のマップを次の点から評価した。

- (1) (Target Curriculum の) シラバスが重複なくトピックごとに配置されうるものであること (どの領域に属しているか判断できること)
- (2) 意味的に近いトピックがマップ上でも近くなっていること (配置に意味があること)
- (3) マップに大きな隙間がないこと (平面が有効に利用されていること)

(1) については、多くの領域に重複がないことで分かる。SE, SP, IS には重複があるが、これは「ソフトウェア工学」, 「社会的視点と情報倫理」, 「インテリジェントシステム」が複合的な内容で重複があることが関係していると考えられる。PF と AL に重複があるのは、「プログラミングの基礎」では基本的な「アルゴリズム」を扱うことが多いことによるものと考えられる。逆に、同じプログラミングに関するものである PF と PL に重複がほとんどないのは、「プログラミングの基礎」では「プログラミング言語」の詳細に入るのを避ける (特にシラバス上) によると考えられる。IM と MR, SP, GV に重複があるが、これは IM の「情報管理」が様々なもののある側面に関するものであるためと考えられる。

*6 1 次元ガウス分布 (正規分布) では $\pm 2\sigma$ には 95.4% の要素が含まれるが、2 次元ガウス分布では $\pm 2\sigma$ には 86.5% の要素が含まれる。

(2) については、マップ上の軸の意味を考察することにする。Isomap により高次元の多様体を展開して 2 次元に射影しているため、軸の意味は大局的に一定というわけではないが、図 6 では、左は社会、右は理論に関するものが多い。また、上は管理、下はハードに関するものが多い。中央は人間に関するものが多い。この解釈のもとで、楕円の領域の幅が狭いものについて考えると、SE は上下の幅が狭いので、シラバスの違いは社会的か理論的かが主で、管理やハードの側面はほとんどないことになる。逆に OS のシラバスは管理とハードの違いがほとんどで社会的か理論的かという側面はほとんどない。AR は左上右下の向きに細長いので、ハードに近いほど理論的という傾向があることになる。これらは領域の性格と合っているものと考えられる。

また、マップの配置に意味があるもう 1 つの傍証としては、教科書の同じ章節がマップ上で近い領域に配置されることがあげられる。実際、6-7 割の章については、その節が同じ領域に配置されている。

(3) については、第 2 象限がやや空気が大きいですが、他については比較的埋められていることが分かる。

4.4 2 学科の情報系カリキュラムの分析

ある大学の分野横断的な A 学科の 3, 4 年生向けの情報系のカリキュラムを Target Curriculum として作成したマップを図 7 に示す。シラバスは主たるトピックの領域の近くに配置されている。関係が深い領域と配置が違うシラバスは図に科目名などを入れてある。

次に、A 学科のカリキュラムを情報学を中心に変更している B 学科の情報系のカリキュラムを Target Curriculum として作成したマップを図 8 に示す。B 学科は A 学科と同じ大学および同じ学部に関する新しい学科である。

分析対象である Target Curriculum の個々のシラバスは、主たるトピックに対応する楕円の中やその近傍に配置されることが期待される。しかし、実際には楕円の領域とは異なる場所にシラバスが配置される場合もありうる。たとえば、A 学科の c04 は、オペレーティングシステムとネットワークに関する基礎的事項を取り扱っており、最も関係がある領域は OS であるが、NC との関係も深く、マップ上

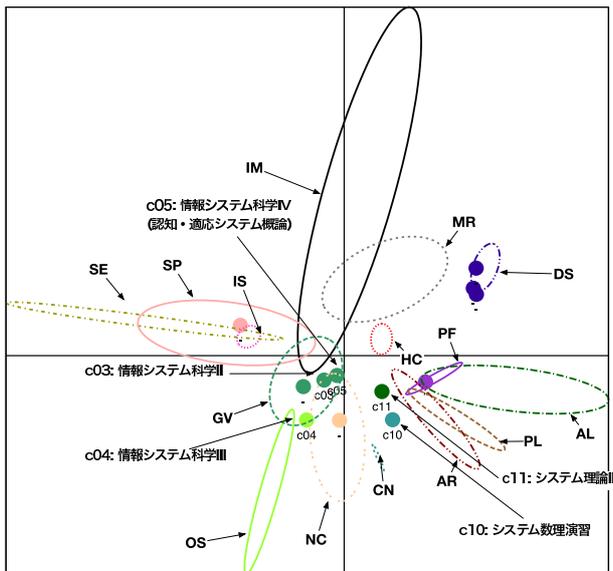


図 7 A 学科 (Target Curriculum) のマップ
Fig. 7 Map of "A" department (Target).

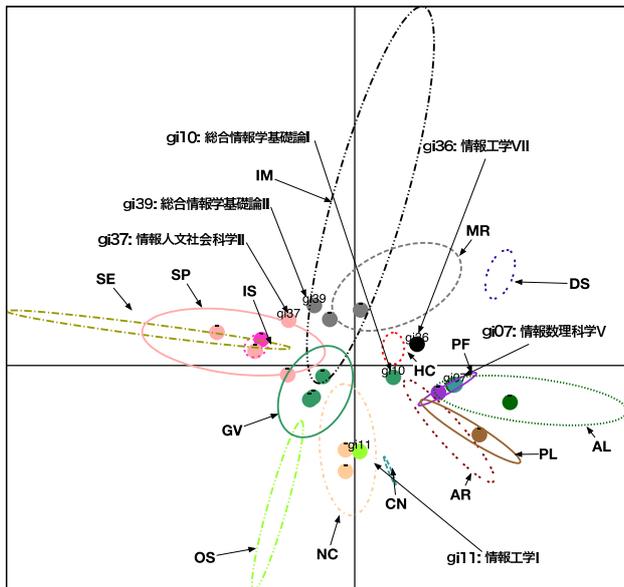


図 8 B 学科 (Target Curriculum) のマップ
Fig. 8 Map of "B" department (Target).

表 2 2 学科の情報系カリキュラムの比較

Table 2 Comparison between CS related curricula of "A" department and those of "B" department.

学科	特徴	シラバス数	最も関係が深い領域の楕円から外れたシラバス数
A	分野横断型	12	3
B	情報学中心型	21	6

DS から AL, PL, NC に移っており、理論 (右) の方で管理からシステム寄り (下向き) のものが増えていることになる。また、MR のあたりと NC のあたりが増えており、メディアとネットワークが強化されていることが分かる。これは横軸では両者とも「人間」に関係するところであり、B 学科で強化した内容と一致している。

(2) については、A 学科で 75%、B 学科で 71% が楕円領域の中にあり、配置から正しく領域が読み取れていることが分かる。領域から外れるシラバスについては次節で検討する。

4.5 不適切なシラバスの検出

A 学科と B 学科のシラバスの中で、想定されている領域から外れているものがある。本節ではそれらを個別に吟味する。

A 学科のシラバスで、想定される領域から外れているものは c04, c10, c11 の 3 つである。c03 は、ソフトウェア工学分野の授業であるが、GV の領域に配置されている。これはシラバスの中に用いられている用語が、Model Curriculum の J07-CS の SE のユニットに含まれる単語とあまり一致しなかったため、学生がどの領域の授業か判断しにくい可能性がある。

c10 は、「簡単な常微分方程式の初期値問題を例として、方程式の厳密解と数値解のそれぞれの性質と、両者の関係について学ぶ。解く方程式は運動方程式、特に単振動の問題とケプラー問題である。」という内容であり、CN との関係が強い (37.1%) が、CN の領域から外れている。これは、Reference Curriculum として用いた教科書に CN に関する内容が少なく、CN の領域が狭すぎることによるものと考えられる。

c11 は、「グラフ、ネットワーク、マトロイド上の組合せ最適化問題とアルゴリズム」という内容であり、AL との関係が強い (24.0%) が AL から外れている。次に関係が強いのは NC (15.2%) であり、それに引かれたためと考えられる。NC と関係が深くなった理由は、シラバス中のグラフ理論の「ネットワーク」がネットワーク技術として解釈されたのではないかと考えられる。用語「ネットワーク」がカリキュラム標準でも多義に使用されているので曖昧さの解消を行うことも考えられるが、学生が誤解ないようにシラバスを書くことも必要であろう。

では NC の領域に配置されている。これをふまえて、A 学科と B 学科のシラバスのマップで以下を評価する。

- (1) Target Curriculum を構成する個々のシラバスが、マップ上にどのように分布するか。
- (2) 主たるトピックに対応する楕円領域から外れたシラバスに注目する。外れたシラバスに対しては、外れた原因を個別に議論する。

(1) については、A 学科のマップと B 学科のマップは 4.3 節の Reference Curriculum を使っているの、その軸の解釈を適用することができる。B 学科のシラバスは A 学科のシラバスより数が多い (表 2) が、カバーされている大きさは同じようなものである。カバーしている領域は、

B 学科のシラバスで、想定される領域から外れているものは gi07, gi10, gi11, gi36, gi37, gi39 の 6 つである。

gi10 は、「総合情報学の各トピックを概要する。」という内容である。GV との関係が最も強い (23.6%) が、AL (16.0%), PL (13.3%), IM (8.7%) などとの関わりもある。シラバスからは、授業ごとに内容が多岐にわたることが読み取れるので、特定の領域に属するものとならない。このような総合的な内容のシラバスについては授業の回ごとにマップするなどが必要になる。

gi11 は、「コンピュータアーキテクチャおよびオペレーティングシステムの基礎を学び、ノイマン型コンピュータが動作する仕組みをハードウェア、ソフトウェアの両面から理解を深める。」という内容である。OS との関係が最も強い (37.0%) が、AR との関係も強い (24.5%)。シラバスからも「コンピュータアーキテクチャ」と「オペレーティングシステム」両方を扱っていることが分かるが、OS と AR の間にある NC の領域に重なることでの誤解が生じる可能性はある。

gi34 は、「問題を解くアルゴリズム (解法)、特にグラフ、ネットワークなどの組合せの対象の上の最適化問題のアルゴリズムの解説とそれを支える線形計画法での双対定理などの数学的理論を学ばせる。」という内容である。CN との関係が最も強い (25.0%) が、DS との関係も強く (22.9%)。AL との関係 (14.8%) もある。シラバスからは DS 的な要素 (グラフ理論) もあることが分かる。したがって、DS と CN の中間にあると解釈できる。

gi36 は、「自然言語 (日本語や英語) を情報処理の対象としてとらえ、その理解や処理をアルゴリズムとデータ構造の観点で理解する。自然言語処理応用技術について、その内容と方式の基本的な理解を得る。」という内容である。IM との関係が最も強い (20.8%) が、AL との関係 (14.7%) や IS との関係 (13.4%) も強い。J07-CS に自然言語処理に関する用語が必ずしも十分ではないため、複数の領域と関係したと考えられる。

gi37 は、「情報通信テクノロジーの発展と社会システムの高度化、求められる職能について理解を深め、高度なソフトウェアとデータベースを用いて実習を行い、非定型的分析能力、非定型的企画立案能力を習得させる。」という内容である。SP との関係が強い (25.9%) が、IM との関係も小さくはない (14.8%)。そのため IM 寄りに配置された。

gi39 は、「情報・知識・メディア・コミュニケーション等の基本概念に関する歴史的・理論的理解を深める。」という内容である。MR との関係が強く (31.6%)、次に関わりが強いのは SP (25.2%) である。これはマルチメディアではなく文系的なメディアの話であるので、情報系で見ると SP 寄りになることは適切である。

まとめると、シラバスの中で領域に関係する用語が少な

く、何に関するシラバスか判断しにくいもの (c03) と多義の用語が使われていて誤解される危険があるもの (c11) では、シラバスの問題点を指摘している。内容が複数の領域にわたるもの (gi10, gi11, gi34, gi37, gi39) は、授業の回ごとにシラバスを分けるなどの工夫をするとより分かりやすくなる可能性がある。J07-CS と関係が強い内容のために J07-CS では分析できないもの (gi36) は、Model Curriculum とのミスマッチの例である。また、教科書にその領域に関係する内容が少ないため領域が狭くなりすぎているもの (c10) は Reference Curriculum とのミスマッチの例である。

4.6 シラバスの改良

図 8 に示した B 学科のカリキュラムを構成する講義 gi07 は、数学的な理論に関する講義である。シラバスの一部を表 3 に示す。本実験では CN と DS の内容に近いとされ、結果としてそれらの領域の中間付近に配置されている。ここで、「gi07 は「DS: 離散構造」の内容を中心としており、マップ上でも DS 領域の近くに配置されるべきである」と考えた場合を例として、4.1 節に述べた作業 B の手順を

表 3 gi07 (情報数理学 V) のシラバス (一部)
Table 3 Syllabus of gi07 (Information mathematical science V).

問題を解くアルゴリズム (解法)、特にグラフ、ネットワークなどの組合せの対象の上の最適化問題のアルゴリズムの解説とそれを支える線形計画法での双対定理などの数学的理論を学ばせる。
計算の複雑さの理論と NP-完全問題の理解。
オートマトンと数理言語学の初歩的理解。チューリング機械と帰納的関数の理論の枠組みの理解。
ネットワークの最短路問題、最大流最小カットの定理から始めて、その下敷きとなる線形計画法の双対定理を取り上げる。一方で多項式時間アルゴリズムで解ける問題と NP-完全問題について理解させる。加えて、オートマトンの理論の初歩を学ばせる。
第 1 回: グラフ理論用語の説明 (双対グラフ, サーキット, カットセットなど) と初歩的知識
第 2 回: グラフの最短路問題の解法 (Dijkstra 法, 行列積算法)
第 3 回: Menger の定理と交代道アルゴリズム
第 4 回: 最大流最小カットの定理
第 5 回: 強欲算法とマトロイド
第 6 回: ポリマトロイド
第 7 回: 線形計画法と双対定理
第 8 回: 2 人零和ゲームの min-max 解
第 9 回: 計算の複雑さの理論, 充足可能性問題
第 10 回: 問題のクラス NP と P, NP-完全問題
第 11 回: 決定性オートマトン
第 12 回: 数理言語学入門
第 13 回: 有限マルコフ連鎖
第 14 回: 吸収確率と平衡分布
第 15 回: チューリング機械と帰納的関数の理論, ゲーデル文と不完全性定理

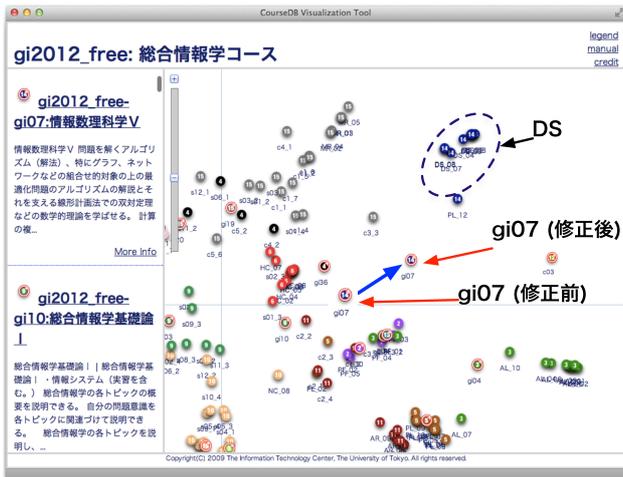


図 9 カリキュラムマップウィンドウ

Fig. 9 The curriculum map window of curriculum analysis tool.



図 10 用語一覧ウィンドウ

Fig. 10 The word list window of curriculum analysis tool.

示す。

図 9 は、本システムの「カリキュラムマップウィンドウ」であり、B 学科のカリキュラムと Reference Curriculum を同時に表示している*7。システムの利用者が DS 関連のシラバスが多く配置されている付近をクリックすると、その近辺のシラバス 3 件と、それらのシラバスで用いられている用語の一覧が「用語一覧ウィンドウ」に表示される (図 10)。

*7 説明の都合上、システムのスクリーンショットに、領域を示す楕円や矢印、領域名、講義名を書き込んである。また、gi07 は修正前後の丸記号を一緒に示している。

利用者が、用語一覧に含まれる用語から、gi07 に関係がありそうな用語を含む「チョムスキー」「正規言語」「文脈自由言語」を「シラバスウィンドウ」でシラバスの中書き加える (図 11) と、マップ上の gi07 の位置は、Reference Curriculum で DS 関連の授業が配置された場所の近くに移動する (図 9)。

このように本システムでは、カリキュラムマップを利用して、シラバスの内容を改良することができる。

5. 結論

本論文では、トピックの生成確率に基づくモデルを用いた LDA と次元圧縮手法である Isomap を組み合わせることで、シラバス文書の集合としてのカリキュラムをマップとして可視化して分析する方法を提案した。その方法に基づいて、J07-CS を Reference Curriculum としたマップを生成し、領域が適切に分離され、軸の意味を読み取ることができた。次に、その Reference Curriculum のマップ上で 2 つの学科の情報系のカリキュラムを比較し、理論系シラバスの傾向の変化を見ることができた。その後、領域の楕円から外れたシラバスを個別に検討することにより、シラバスの曖昧さ、領域 CN の Reference Curriculum のシラバスの不足、複数の領域にまたがるシラバスの配置の問題が明らかになった。

この Reference Curriculum のように基準となるマップを生成したうえで、比較対象となる文書群—本研究においては比較対象のカリキュラムに相当するシラバス文書の集合—を射影する方法は、文書群からマップする岩田ら [11] の方法ではできないし、筆者らの知る限りではそのほかにも見当たらない。

なお、複数の領域にまたがるシラバスの配置の問題については、そのようなシラバスの表示方法についても検討していきたい。また、マップでは、J07-CS のマップ (図 6) の AR と OS、CN と DS のように、異なる分野のシラバスがマップ上の狭い範囲に固まることがある。これは、LDA の出力である 15 次元のトピック空間が、2 次元のマップに完全には展開できず、一部が重なった状態になっているためと考えられる。これを解決するために、3 次元以上の配置を様々な角度から見られるようにするなどの方法を検討していきたい。

本研究で分析した 2 つの学科はいずれも分野横断的な学科であることから、情報系のカリキュラムに関しても比較的広い範囲の内容を含んでおり、J07-CS を用いた分析対象として適切であると筆者らは考えている。同様の情報系のカリキュラムを提供する学科であっても、理論的な講義や実践的な講義に偏り、J07-CS の一部の領域の内容しか扱わない学科もあるからである。しかし、今後は他の大学や学科のデータを用いたり、カリキュラム設計に携わる教職員に実際に使ってもらったりすることで、本手法の有効



図 11 シラバスウィンドウ

Fig. 11 The syllabus window of curriculum analysis tool.

性を検証する必要があると考えている。また、本手法自体は、基準となるカリキュラムがあれば他分野のカリキュラムの分析に適用することも可能であるので、今後、他分野のカリキュラムの分析も進めていきたい。

謝辞 大変有益なご指摘をいただいた査読者の方々に深く感謝いたします。

参考文献

[1] Tucker, A.B., Barnes, B.H., et al.: Computing Curricula 1991 Report of the ACM/IEEE-CS Joint Curriculum Task Force (1990).
 [2] Asuncion, A. and Newman, D.J.: UCI Machine Learning Repository (2007).
 [3] Bishop, C.M.: *Pattern Recognition and Machine Learning*, Springer, New York, NY, USA (2006).
 [4] Blei, D.M.: LDA-C (2006), available from <http://www.cs.princeton.edu/~blei/lda-c/>.
 [5] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.993-1022 (2003).
 [6] Carvalho, G.S. and Clément, P.: Construction and validation of the instruments to compare teachers' conceptions and school textbooks of 19 countries: The European Biohead-Citizen project (2007).
 [7] Griffiths, T.L., Steyvers, M. and Tenenbaum, J.B.: Topics in Semantic Representation, *Psychological Review*, Vol.114, No.2, pp.211-244 (2007).
 [8] Harvard Graduate School of Education: The Collaborative Curriculum Design Tool (CCDT) (2010), available from <http://learnweb.harvard.edu/ccdt/> (accessed 2010-02-22).
 [9] Ida, M., Nozawa, T., Yoshikane, F., Miyazaki, K.

and Kita, H.: Syllabus database and Web service on higher education, *7th International Conference on Advanced Communication Technology (ICACT2005)*, Vol.1, pp.415-418 (2005).
 [10] 井田正明, 野澤孝之, 芳鐘冬樹, 宮崎和光, 喜多一: シラバスデータベースシステムの構築と専門教育課程の比較分析への応用, *大学評価・学位研究*, No.2, pp.85-97 (2005).
 [11] 岩田具治, 山田武士, 上田修功: トピックモデルに基づく文書群の可視化, *情報処理学会論文誌*, Vol.50, No.6, pp.1649-1659 (2009).
 [12] Mima, H.: MIMA search: A structuring knowledge system towards innovation for engineering education, *Proc. COLING/ACL on Interactive Presentation Sessions*, Morristown, NJ, USA, Association for Computational Linguistics, pp.21-24 (2006).
 [13] Ota, S. and Mima, H.: Machine Learning-based Syllabus Classification toward Automatic Organization of Issue-oriented Interdisciplinary Curricula, *Procedia - Social and Behavioral Sciences*, Vol.27, pp.241-247 (2011).
 [14] Pedroni, M., Oriol, M. and Meyer, B.: A framework for describing and comparing courses and curricula, *Proc. 12th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, ITiCSE '07*, New York, NY, USA, ACM, pp.131-135 (2007).
 [15] Pinar, W.F., Reynolds, W.M., Slattery, P. and Taubman, P.M.: *Understanding Curriculum: An Introduction to the Study of Historical and Contemporary Curriculum Discourses*, Peter Lang Pub. Inc. (1995).
 [16] Ronchetti, M. and Sant, J.: Towards automatic syllabi matching, *ITiCSE '09: Proc. 14th Annual ACM SIGCSE Conference on Innovation and Technology in Computer Science Education*, New York, NY, USA, ACM, p.379 (2009).

- [17] Roweis, S.T. and Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding, *Science*, Vol.290, No.5500, pp.2323-2326 (2000).
- [18] Sekiya, T., Matsuda, Y. and Yamaguchi, K.: Analysis of Computer Science Related Curriculum on LDA and Isomap, *ITiCSE'10, Proc. 15th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education*, pp.48-52 (2010).
- [19] Sekiya, T., Matsuda, Y. and Yamaguchi, K.: Development of a Curriculum Analysis Tool, *ITHET 2010, 9th International Conference on Information Technology Based Higher Education and Training* (2010).
- [20] Tenenbaum, J.B., de Silva, V. and Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction, *Science*, Vol.290, No.5500, pp.2319-2323 (2000).
- [21] The Joint Task Force on Computing Curricula IEEE Computer Society/Association for Computing Machinery: Computing Curricula 2001 Computer Science (2001), available from (http://www.acm.org/education/curric_vols/cc2001.pdf/view).
- [22] Tungare, M., Yu, X., Cameron, W., Teng, G., Perez-Quinones, M.A., Cassel, L., Fan, W. and Fox, E.A.: Towards a syllabus repository for computer science courses, *Proc. 38th SIGCSE Technical Symposium on Computer Science Education*, pp.55-59 (2007).
- [23] Wiske, M.S., Sick, M. and Wirsig, S.: New technologies to support teaching for understanding, *International Journal of Educational Research*, Vol.35, No.5, pp.483-501 (2001).
- [24] 情報処理学会コンピュータ科学教育委員会：カリキュラム標準コンピュータ科学 J07-CS 報告書 (2009年1月20日) (2009), 入手先 (http://www.ipsj.or.jp/12kyoiku/J07/20090407/J07.Report-200902/4/J07-CS_report-20090120.pdf).



山口 和紀 (正会員)

1979年東京大学理学部数学科卒業。1981年東京大学理学部助手。1985年理学博士(東京大学)。1989年筑波大学電子情報工学系講師。1992年東京大学教養学部助教授。1999年東京大学情報基盤センター教授。2007年より東京大学大学院総合文化研究科教授。コンピュータのためのモデリング全般に興味を持つ。



関谷 貴之 (正会員)

1972年生。1995年東京大学工学部精密機械工学科卒業。2000年東京大学大学院工学系研究科精密機械工学専攻博士課程単位取得退学。同年東京大学情報基盤センター助手。2007年より同センター助教。博士(工学)。教育

支援システムに関する研究に従事。



松田 源立 (正会員)

1974年生。1997年東京大学教養学部基礎科学科第二卒業。2002年東京大学大学院総合文化研究科広域科学専攻広域システム科学系博士課程修了。博士(学術)。学術振興会特別研究員(PD)を経て、2006年青山学院大学理

工学部情報テクノロジー学科助手。2008年同助教。ニューラルネット、機械学習、信号処理等の研究に従事。