The 35th Annual Conference of the Japanese Society for Artificial Intelligence, 2021

# 半教師ありNMFを用いた専門分野と講義の関係推定

Estimation of Relationships Between Field of Research and Classes by Using Semi-Supervised NMF

山本 京佳 *1　山西 良典 *1　松下 光範 *1
Kyoka Yamamoto　Ryosuke Yamanishi　Mitsunori Matsushita

*1関西大学総合情報学部
Faculty of Informatics, Kansai University

The goal of this study is to visualize the relationships between classes and specialty (i.e., laboratory) for supporting students to determine the classes with the future direction of study. The students choose their classes by themselves because the university curriculum is highly flexible. Each class should be connected to some specialties though, it is difficult for students to understand the relationships from just syllabus without sufficient knowledge. This paper proposes a method to estimate the relationships between classes and laboratories in the faculty. The proposed method applies semi-supervised non-negative matrix factorization to reveal the common factors of knowledge in each combination of laboratory and class. It was suggested that reasonable results for the relationship between the laboratory and classes were calculated by the proposed method. We believe that it is possible for students to understand which class should relate to which laboratory.

## 1. はじめに

大学のカリキュラム構成は自由度が高く，学生は自身の学びたい専門分野に沿った講義を選択することが求められる．そのため，学生は履修時に講義計画が記載されたシラバスを参照し，自身の学びたい専門分野と講義で扱う知識のつながりを把握しながら，選択するべき講義を決定する必要がある．しかし，専門的な学問についての知識が少ない学生にとって，シラバスから専門分野と講義で扱う知識のつながりを把握することは容易ではない．

大学での受講デザインに着目した研究では，講義の内容を要約した情報を含むシラバスを用いることで，講義間の関係推定が行われている．由谷ら [由谷 06] は，シラバスから抽出した専門用語の重みを用いて講義間の類似度を計算することで，複数の講義間の関係を分析している．野澤ら [宮崎 05] は，シラバスを用いて講義をクラスタリングし，各クラスタの特徴語による意味付けおよび各カリキュラムでの講義のクラスタ分布に着目したカリキュラム間の比較を行っている．美馬ら [Mima 06] は，シラバスのテキストを解析して講義間の類似度を抽出することで科目間関係図を可視化する MIMA search を提案し，カリキュラム全体像の俯瞰と講義間の関係把握を可能にしている．岡田ら [岡田 17] は，学習の順序を設定した学習パスに沿った「学ぶべきもの」の提供によって自立学習が促進されると提唱し，講義間の関係を学習者の情報とともに可視化するコンセプトマップを作成している．しかしながら，これらの研究では主に講義間の関係に着目しており，講義で学んだ知識や技術が専門分野（すなわち，研究室活動）においてどのように生かされるのかについては扱われていない．

本研究では，学生の専門分野を見据えた講義選択が可能になることを目指し，専門分野と講義の関係を可視化するための推定手法を提案する．提案手法では，因子分解技術を応用して専門分野と講義の共通因子を顕在化することで，専門分野と講義の関係を説明可能にする．

連絡先: 山本京佳，関西大学総合情報学部，〒 569-1095, 大阪府高槻市霊仙寺町 2–1–1, Tel: 072-690-2437, Fax: 072-690-2491, {k166981, ryama, m_mat}@kansai-u.ac.jp

## 2. 提案手法

提案手法では，各専門分野での研究内容が把握可能な論文情報と各講義情報が得られるシラバスを対象として，因子分解技術を適用することで，専門分野と講義の共通因子を顕在化する．因子分解の基本的な考え方は，以下のように表現される．

$$\text{観測変数} = \text{共通因子} \times \text{独自因子}.$$

このとき，異なる観測変数に共通して見られる特徴を共通因子，共通因子では説明できない各観測変数固有の要素を独自因子とする．本稿では，因子分解技術の中でも，非負値行列因子分解（Non-negative Matrix Factorization：以下，NMF）[Lee 01] を拡張して，あらかじめ教師情報を与えた上で因子分解を行う半教師あり NMF（Semi Supervised NMF：以下，SSNMF）[Lee 10] を用いる．

### 2.1 NMF

NMF における因子分解は，式 (1) によって定義される．観測変数行列，基底行列，アクティベーション行列をそれぞれ $\boldsymbol{Y}$，$\boldsymbol{H}$，$\boldsymbol{U}$ としたとき，与えられた $\boldsymbol{Y}$ を $\boldsymbol{H}$ と $\boldsymbol{U}$ の積によって近似する．

$$y_{ij} \approx y'_{ij} = \sum_{k=1}^{K} h_{ik} u_{kj}, \tag{1}$$

ここで，$y$，$h$，$u$ はそれぞれ $\boldsymbol{Y}$，$\boldsymbol{H}$，$\boldsymbol{U}$ の各要素，$i$ と $j$ は行列内の要素のインデックスを示し，$K$ は基底数を示す．

行列 $\boldsymbol{Y}$ と行列 $\boldsymbol{HU}$ の誤差を最小化することで，行列 $\boldsymbol{HU}$ を更新していく．行列 $\boldsymbol{Y}$ と行列 $\boldsymbol{HU}$ の誤差最小化で用いられる誤差関数は，いくつかあるが，本稿では，下式で表される二乗 Euclid 距離 $D_{Euclid}$ を用いた．

$$D_{Euclid}|(\boldsymbol{Y}, \boldsymbol{HU})| = ||\boldsymbol{Y} - \boldsymbol{HU}||^2. \tag{2}$$

### 2.2 SSNMF

半教師あり NMF は基底ベクトルとしてあらかじめ用意した教師ベクトルを与えて因子分解を行う手法である．SSNMF は，音源分離で利用される事例 [林 16, 北村 13] が多く報告さ

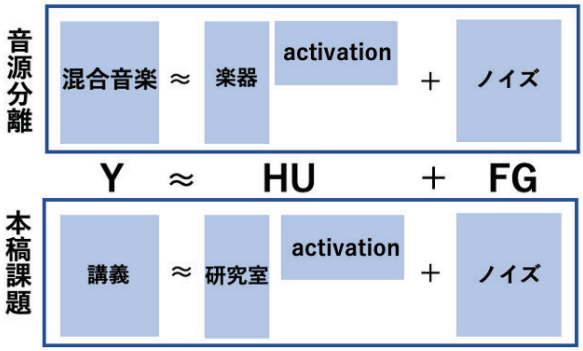



図 1: 音源分離と本稿で扱う課題の SSNMF 適用の対応


図 2: Bag of Words 法の適用によって得られた研究室と講義を示す数値行列

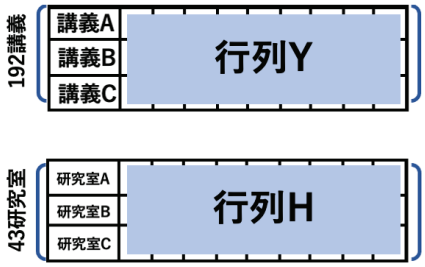



図 3: 本稿における SSNMF の行列 Y と行列 H

れており，これらの研究では音源のスペクトル構造を教師ベクトルとして与えることで各音源のアクティベーションを獲得し，音源からの自動採譜などに応用されている．SSNMF は下式で定義される．

$$\boldsymbol{Y} \approx \boldsymbol{HU} + \boldsymbol{FG}, \tag{3}$$

ここで，$\boldsymbol{H}$ と $\boldsymbol{U}$ は教師ベクトルとそのアクティベーション行列，$\boldsymbol{FG}$ はノイズ項を示す．このとき，$\boldsymbol{Y}$，$\boldsymbol{H}$，$\boldsymbol{U}$，$\boldsymbol{F}$，$\boldsymbol{G}$ はそれぞれ，$S \times N$，$L \times N$，$L \times S$，$N \times R$，$R \times L$ で定義されるサイズの行列を示す．$L$ と $R$ はそれぞれ，教師ベクトルとノイズ項のベクトル基底数を示す．

図 1 に，音源分離における SSNMF と本稿における SSNMF の適用の対応を示す．SSNMF が音源分離で利用される場合には，複数の音源が多重に混合された音声を示す周波数と時刻で構成される行列 $\boldsymbol{Y}$ を，抽出したい楽器の周波数構造を示す教師ベクトル $\boldsymbol{H}$ で分解することにより，多重に音源が混合された音楽における楽器の各時刻の占有率（activation）を示す行列 $\boldsymbol{U}$ が抽出される．本稿では，専門分野と講義の関係推定を音源分離のしくみに当てはめる．つまり，様々な知識が多重に混合された講義を示す行列 $\boldsymbol{Y}$ を，専門分野で扱う知識を示す行列 $\boldsymbol{H}$ で分解することにより，専門分野で扱う知識を講義の関係を示すアクティベーション行列 $\boldsymbol{U}$ として求めることができる．本稿では，専門分野と講義の情報リソースとして，非負値で表現可能な情報を扱う．

## 3. 提案手法による専門分野と講義の関係推定

提案手法の専門分野と講義の関係推定における有用性を検証した．以下，使用したデータと提案手法におけるパラメータ設定等について説明する．

### 3.1 データの準備

本稿では，文理融合学部であり，様々な専門性を一学部内で横断的に学ぶことができる関西大学総合情報学部（以下，SJ 学部）を推定対象とした．SJ 学部では，プログラミングやアルゴリズムをはじめとする情報学の基礎理論（以下，C 系），情報メディアやコミュニケーションにおける情報処理（以下，M 系），経営，経済，心理，政治などの各分野における情報処理（以下，S 系），といった様々な専門分野についての講義が用意されている．また，各研究室では学部に所属する各教員がそれぞれの専門性に関する研究指導を行っている．以下，SJ 学部についての専門分野と講義をデータとして取得するための方法を示す．

#### 3.1.1 専門分野の情報

一つの研究室を一つの専門分野として扱い，専門分野についての情報リソースとして 2019 年度 SJ 学部卒論概要集から取得できた 43 研究室の卒業論文概要を対象とした．卒業論文概要から，研究室ごとに指導教員名，学籍番号・名前，参考文献を除いた本文全てを抽出し，専門分野で扱う知識として使用した．

#### 3.1.2 講義の情報

講義情報のリソースとして，大学のホームページから取得した 2020 年度 SJ 学部のシラバスを用いた．ここで，推定対象の講義は 2020 年度に開講されたもののうち，外国語科目やスポーツ実習を除いた 192 科目とした．

シラバスには，科目名，授業形態，担任者名，授業概要，授業計画，到達目標，授業方法，成績評価等が記載されている．本稿では，このうち，授業内容を示すと考えられる授業概要，授業計画，到達目標に記載された文章をデータとして扱った．

#### 3.1.3 テキストの正規化

3.1.2 節と 3.1.1 節でそれぞれ取得したシラバスと卒論概要のテキストデータに対して，半角英数字および記号の正規化を行い，改行，スペースを取り除いた．その後，形態素解析エンジン mecab-python3（ver.1.0.1）を用いて単語ごとに分割し [Kudo 04]，名詞のみを抽出した．単語分かち書き辞書には mecab-ipadic-NEologd[*1] を選択した．その際，ストップワードに設定した単語を除外した [*2]．

### 3.2 データの数値表現化と分散表現化

3.1 節で得られたデータに対して，Bag of Words（BoW）法を使用して数値表現に変換した．ここで，図 2 のように研

*1 https://github.com/neologd/mecab-ipadic-neologd
*2 http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/Japanese.txt

表 1: 考察対象の研究室とその専門分野を示す研究概要．研究概要については主催する研究者の研究紹介から抜粋．

| 研究室 | 研究概要 |
|---|---|
| a 研究室 | ロボットを使った認知機能の研究 |
| b 研究室 | 情報セキュリティとセキュアネットワークに関する研究開発 |
| c 研究室 | インタラクションデザイン： 使いやすいシステムの実現を目指して |
| d 研究室 | ヒューマン–メディア<br>インタラクションとコミュニケーションの新しいデザインの追求<br>擬人化メディア（ロボット/仮想エージェント等）<br>環境型メディア（音楽・音声・映像・SNS・モバイルシステム）等を含む |
| e 研究室 | メディアをつくる、社会をとらえなおす |
| f 研究室 | コンピュータグラフィックスとヒューマンインタフェース技術に関する研究 |
| g 研究室 | フィールドワークから学ぶマーケティング |
| h 研究室 | 政治データベースの構築と政治データ分析 |
| i 研究室 | 経営戦略，技術経営，国際経営 |

究室と講義を結合した数値行列を作成した．その際，単語の重み付けのため出現頻度が同文書あたり 1 未満の単語を削除し，研究室と講義を 7560 単語の出現頻度によって数値表現した．得られた 235（43 研究室＋ 192 講義）× 7560（BoW 単語数）の行列に対して，NMF を適用することで 500 次元の分散表現を獲得した．

### 3.3 SSNMF の適用

3.2 節で得られた行列に対して，SSNMF を適用した．図 3 に示すように，研究室と講義の集合を示す 2 種類のベクトル集合へ分割して，$\boldsymbol{Y}$ と $\boldsymbol{H}$ を取得した．この 2 種類の行列集合は，どちらも 1) 非負値，2) 行列，3) 各集合の知識内容を表現可能という 3 つの条件を満たしている．

SSNMF の適用では，行列 $\boldsymbol{Y}$ に講義情報を，教師ベクトルである行列 $\boldsymbol{H}$ に専門分野を与えた．ここで，SSNMF の近似におけるイテレーション数は実験的に 80,000 回とした．

## 4. 結果と考察

本稿では，提案手法の適用によって得られたアクティベーション行列の中から，表 1 に示す異なる専門分野の 9 研究室について考察する．提案手法の適用によって，図 4 に示すように各研究室に対してどのような講義がどれほど v するのか可視化された．表 2 に，各研究室のアクティベーション値上位 4 位までの講義を示す．

### 4.1 研究室の専門分野とアクティブな講義の妥当性

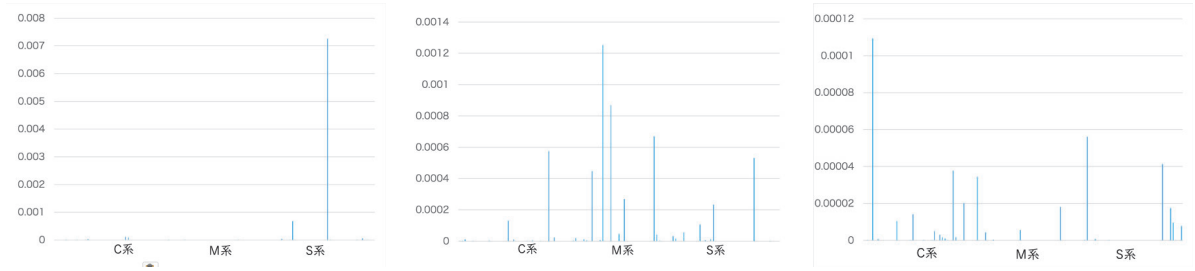
C 系の専門分野を扱う研究室では，良好な結果が得られた．ロボットを使った認知機能の研究を行っている a 研究室で最もアクティベーション値が高いものとして,「機械学習実習」が挙げられている．ロボットは動作と多様な状況に対して対応するため機械学習の適用が望まれており，妥当な結果が得られたと考える．セキュアネットワークに関する研究開発を行っている b 研究室では,「ネットワーク実習」が高いアクティベーション値を示している．また，インタラクションデザインに関する研究を実施している c 研究室では，ハードウェアからソフトウェアまであらゆる基礎知識が必要であるが,「ハードウェアアーキテクチャ」や「コンピューティングの言語」などが関係の深い講義として検出された．また，C 系の専門分野の中でもエージェントを用いてコミュニケーション表現についての研究を実施している d 研究室では,「認知的人工物のデザイン」「ヒューマンエージェントインタラクション」が関係深いことを示している．計算機科学以外の社会や芸術と関係の深い研究を専門分野とする研究室においても妥当な結果が得られたと考えられる．ヒューマンインタフェースにおける情報の提示方法のなかでも，物理学や工学分野における可視化システムなどを研究している f 研究室では「コンピュータの物理」が関係深いことを示している．情報メディアの変化と日常生活の関わりについてを研究する e 研究室では,「地域メディア論」や「映像メディアと現代社会」などが高いアクティベーション値を示している．コンピュータグラフィックスの中でも，特に視覚や色彩，錯視を対象とする g 研究室では,「デザイン論」や「感性情報処理」の講義が深く関係すると示された．一方で，本来は関係が薄いと考えられる「経済政策論」や「ミクロ政治データ分析実習」などについても高いアクティベーション値が示されており，原因の分析と対策についてはさらなる議論が必要となる．政治データの分析を行っている h 研究室では，政治についての講義が上位 4 講義として示された．また，経営についての研究を行っている i 研究室では,「経営情報システム論」「経営学」の講義が高いアクティベーション値を示し，研究室の専門分野に関係が深い講義が提案手法によって検出されたと考察される．

表 2: 各研究室のアクティベーション値が高い上位 4 講義

| 研究室 | 講義 | activation |
|---|---|---|
| a 研究室 | 機械学習実習 | 0.000344977 |
| | ＣＧ実習（３Ｄコンテンツ開発） | 0.000335494 |
| | オブジェクト指向プログラミング（Ｊａｖａ） | 0.000201372 |
| | モバイル・コンピューティング | 0.000159336 |
| b 研究室 | 機械学習実習 | 0.001347930 |
| | ネットワーク実習 | 0.000697389 |
| | プログラミング実習（C） | 0.000129344 |
| | プログラミング入門 | 0.000109343 |
| c 研究室 | ハードウェアアーキテクチャ | 0.001920421 |
| | コンピュータの言語 | 0.001217381 |
| | 情報デザイン | 0.000639332 |
| | 機械学習実習 | 0.000613044 |
| d 研究室 | テーマ別研究（認知的人工物のデザイン） | 0.000707536 |
| | ヒューマンエージェントインタラクション | 0.000620781 |
| | 環境経済学 | 0.000473427 |
| | コミュニケーションと行為 | 0.000315444 |
| e 研究室 | 地域メディア論 | 0.001252894 |
| | 経営学 | 0.001031541 |
| | 異文化コミュニケーション | 0.000867930 |
| | 映像メディアと現代社会 | 0.000670041 |
| f 研究室 | 制作実習（地域コンテンツ） | 0.000316359 |
| | コンピュータの物理 | 0.000295495 |
| | エンターテインメント論 | 0.000210567 |
| | マクロ政治データ分析実習 | 6.64042E-05 |
| g 研究室 | デザイン論 | 0.000924566 |
| | 経済政策論 | 0.000607605 |
| | 感性情報処理 | 0.000497757 |
| | ミクロ政治データ分析実習 | 0.000488301 |
| h 研究室 | マクロ政治データ分析実習 | 0.007894836 |
| | ミクロ政治分析 | 0.007254449 |
| | 政治学 | 0.004563811 |
| | 政治過程論 | 0.000678525 |
| i 研究室 | 経営情報システム論 | 0.005897955 |
| | アルゴリズム解析・設計 | 0.001550269 |
| | 経営学 | 0.001145948 |
| | 人工知能 | 0.000677252 |

### 4.2 提案手法の問題点の分析と展望

提案手法の問題点を x 研究室を取り上げて，分析する．情報セキュリティネットワークを専門分野とする x 研究室では，表 3 に示すように情報セキュリティネットワークとは関係が




(a) 政治データベースの構築と政治データ分析を行う h 研究室

(b) 情報メディアと社会の関係分析に関する研究を行う e 研究室

(c) 情報セキュリティとセキュアネットワークに関する研究開発を行う b 研究室

図 4: 研究室ごとに得られる講義アクティベーションの例．各図中で左から順に，3.1 節で示した C 系，M 系，S 系の順で講義がソートされており，それぞれの講義に対して各研究室がどれほど関係するのかが縦軸の値から読み取れる．

表 3: X 研究室での値の高かった上位 10 講義

| 講義 | activation |
|---|---|
| 社会心理学 | 0.00333296 |
| ネットワーク実習 | 0.00017812 |
| 情報システムの基礎 | 0.00003348 |
| 質的調査法 | 0.00002367 |
| 情報セキュリティ論 | 0.00002014 |
| エンターテインメント・コンピューティング | 0.00001032 |
| 認知科学 | 9.7814E-06 |
| Ｗｅｂ情報システム論 | 6.4174E-06 |
| 地球観測の情報処理 | 5.8612E-06 |

薄いと考えられる「社会心理学」が極端に高いアクティベーション値を示した．この原因として，言葉の多義性が考えられる．X 研究室の卒業論文概要では，ネットワーク分野におけるサーバ攻撃に用いられる「攻撃」という単語が取り上げられていた．一方で,「社会心理学」の講義では，社会心理分野における人に対する「攻撃」という単語が取り上げられている．どちらも,「攻撃」という他の研究室や講義では使われることが少ない単語が用いられており，これらの研究室と講義を結びつける特徴的な単語であると誤判断されたと推察される．最近では，YouTube のチェスに関するチャンネルが，コメント欄の情報から人種差別を扱うチャンネルと誤認識されたと分析する研究 [Sarkar 21] も報告されており，自然言語処理において共通の課題であると考えられる．

今後の課題としては，アクティベーション値の高い講義が専門分野を学ぶうえで履修すべき講義か，アクティベーション値の低い講義の中に専門分野を学ぶうえで履修すべき講義が含まれていないかを研究者へのインタビューなども含めて詳細な検証を行う必要がある．また，SJ 学部以外の他学部や他大学でも同様の結果が得られるかについても検証する必要がある．

## 5. おわりに

本稿では，専門分野と講義の関係を因子分解技術を用いて推定し，可視化する手法を提案した．提案手法の適用により，SSNMF を用いて研究室と講義で扱う知識の共通因子を顕在化させた．出力結果を研究室ごとに考察したところ，研究室と講義で扱う知識の関係について概ね妥当な結果が得られたと示唆され，専門分野と講義で扱う知識の関係をアクティベーションの値から読み取れることを確認した．今後は，専門分野と講義の数値表現に単語の局所的なコンテキスト情報を考慮可能な分散表現を適用するほか，他学部への適用や企業情報などとの関係推定などにも応用し，それぞれの有用性を検討する．

## 参考文献

[Kudo 04] Kudo, T., Yamamoto, K., and Matsumoto, Y.: Applying conditional random fields to Japanese morphological analysis, in *Proc. 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237 (2004)

[Lee 01] Lee, D. D. and Seung, H. S.: Algorithms For Non-negative Matrix Factorization, in *Advances In Neural Information Processing Systems*, pp. 556–562 (2001)

[Lee 10] Lee, H., Yoo, J., and Choi, S.: Semi-Supervised Nonnegative Matrix Factorization, *IEEE Signal Processing Letters*, Vol. 17, No. 1, pp. 4–7 (2010)

[Mima 06] Mima, H.: MIMA search: a structuring knowledge system towards innovation for engineering education, in *Proc. COLING/ACL 2006 Interactive Presentation Sessions*, pp. 21–24 (2006)

[Sarkar 21] Sarkar, R. and Khudabukhsh, A. R.: Are Chess Discussions Racist? An Adversarial Hate Speech Data Set, in *Proc. 35th AAAI Conference on Artificial Intelligence*, SA-379 (2021)

[岡田 17] 岡田 卓弥, 吉川 雅修, 岩沼 宏治：学習者の情報とシラバスを用いたコンセプトマップによる自律学習支援, 第 31 回人工知能学会全国大会論文集, 3N2-4in2 (2017)

[北村 13] 北村 大地, 猿渡 洋, 鹿野 清宏, 近藤 多伸, 高橋 祐：基底変形型教師あり NMF による実楽器信号分離, 電子情報通信学会技術研究報告, Vol. 112, No. 388, pp. 13–18 (2013)

[林 16] 林 亜紀, 亀岡 弘和, 松林 達史, 澤田 宏：非負値周期成分分析手法による音楽音響信号の音源分離, 日本音響学会 2016 年春季研究発表会講演論文集, pp. 639–642 (2016)

[宮崎 05] 宮崎 和光, 井田 正明, 芳鐘 冬樹, 野澤 孝之, 喜多 一：電子化されたシラバスに基づく学位授与事業のための科目分類支援システムの試作, 情報処理学会論文誌, Vol. 46, No. 3, pp. 782–791 (2005)

[由谷 06] 由谷 真之, 森 幹彦, 喜多 一：電子シラバスを用いた大学教養教育における科目選択支援, 第 68 回情報処理学会全国大会講演論文集, No. 1, pp. 465–466 (2006)