卒業研究

「LDAトピックモデルと履修成績を用いた履修レコメンドシステム」

国際文化学部　国際文化学科

1686592c 宮崎仁弥

指導教員：村尾　元教授

副指導教員：康　敏教授

目次

1. はじめに

1-1.　研究の背景と目的

本論文では「履修成績を用いて授業をおすすめするシステムの開発」について記す。現在大学生が受講できる授業の数は大変多くなっている。例えば、神戸大学国際人間科学部の2020年に開講された授業数は約1500件である。そのため、学生は自分の趣味・嗜好に合わせて履修することが可能になっており、授業選択の自由度が高くなっている。しかし授業が多様化した反面、履修計画を建てることは煩雑化した。数ある科目の中からシラバスを確認し、自分が興味を持てる授業なのかなどの判断をしながら履修する科目を探し出すことはなかなか時間がかかる。神戸大学の履修神戸大学国際人間科学部グローバル文化学科の学生の履修科目とその成績のデータをもとに、科目選択の効率化や自分の知らなかった得意・興味のある科目の発見を促すことが本研究の目的である。

1-2.　本論文の構成

本論文の構成は次のようになっている。第2章では、先行研究を紹介する。第3章では、本研究で使用した技術について説明する。第4章では、今回用いた手法について述べる。第5章では、レコメンドシステムの評価の方法とその結果について述べる。第6章ではまとめと今後の課題・展望について述べる。

2. 関連研究

本研究と同様に大学生を対象に科目を推薦するシステムに関する既存研究が存在する。竹森ら[1]は、学部新入生を対象に教養科目を推薦するシステムの設計を行った。各科目の特徴を doc2vec を用いてベクトル化し、その科目ベクトルに対してウォード法を用いたクラスタリングを行った。その次に高校主要科目の 5 科目ごとに科目ベクトルを作成する。各クラスタに属する各科目に対し、高校科目ベクトルとの類似度を計算し 0~5 の値で正規化し、ラーダーチャートに表し可視化する。大学の科目のクラスタのキーワードを可視化したワードクラウドを見て、学生はクラスタを選択し、そのクラスタの中の科目から学生はレーダーチャートも参考にしながら科目を選択する。科目をクラスタリングしているという点で共通しているが、本研究では成績を用いているという点で異なる。

西森ら[2]は TF-IDF とコサイン類似度を用いて科目間の類似度を求めた。履修する科目と履修済みの科目の類似度に直接 GPA をかけることで、科目の成績を推定する。履修済みの科目に対して、推定を行ったところ、無作為に推定した場合より絶対平均誤差が低いことを明らかにした。本研究では類似度に直接 GPA を掛け合わせるのではなく、トピックベクトルに GPA を掛け合わせている

3. 使用した技術

3-1.　TF-IDF

TF-IDF は、文書内に出現する単語について TF（出現頻度）と IDF（逆文書頻度）からその単語の重要度を求める手法である。TF とは Term Frequency の略である。これは各文書での単語の出現頻度を意味する。関数 f を出現頻度を求める関数とし、文書d_jにおける単語t_iの出現頻度を表したものが以下の式である。

$$\mathrm{tf}(\mathrm{t_i},\ \mathrm{d_j}) = \frac{\text{文書}\mathrm{d_j}\text{内の単語}\mathrm{t_i}\text{の出現回数}}{\text{文書}\mathrm{d_j}\text{内の全ての単語の出現回数の和}} = \frac{\mathrm{f}(\mathrm{t_i},\ \mathrm{d_j})}{\sum_{\mathrm{t_k}\in\mathrm{d_j}} \mathrm{f}(\mathrm{t_k},\ \mathrm{d_j})}$$

しかし、TF のみではどの文書にも現れる単語の値も大きくなってしまう。そのような単語は文書の特徴を表しているとは考えにくい。そこで用いられるのが IDF である。IDF は Inverse Document Frequency の略である。これはある単語が含まれる文書の割合の逆数を表す。その単語の出現する文書の数が少ないほどこの値は大きくなる。ある文書集合における単語について考える場合、を単語が出現する文書数とすると、IDF 値は以下の式(1)から求められる。

$$\mathrm{idf}(\mathrm{t_i}) = \log\left(\frac{\text{総文書数}}{\text{単語}\mathrm{t_i}\text{が出現する文書数}}\right) = \log\left(\frac{\mathrm{N}}{\mathrm{df}(\mathrm{t_i})}\right) \quad (1)$$

TF-IDF は TF 値と IDF 値を掛け合わせる以下の式(2)で求められる。

$$\mathrm{tfidf}(\mathrm{t_i}, \mathrm{d_j}) = \mathrm{tf}(\mathrm{t_i}, \mathrm{d_j})\mathrm{idf}(\mathrm{t_i}) \quad (2)$$

それにより、ある文書での出現回数は多いが、他の文書にはあまり出現しない単語の TF-IDF 値は大きくなる。TF-IDF 値が大きい単語ほどその文書の特徴を表していると言える。

3-2. LDA トピックモデル

LDA トピックモデルは、文書の確率的生成モデルとして提案された。LDA では一つの文書に複数のトピックが存在すると仮定し、そのトピックの分布を離散分布としてモデル化する[3]。本研究ではシラバスの各授業のトピックの分布を LDA を使って求めている。表 1 は科目とそのトピック分布の一例である。現代 IT 入門 A はトピック 5 の値が一番大きいため、トピック 5 に最も属していると考えられ、グローバル社会動態発展演習 A はトピック 4 以外の値が 0 であるため、トピック 4 にのみ属していると考えられる。本研究ではこのトピック分布をトピックベクトルとみなし計算を行う。

表 1. 科目とトピック分布の一例

科目名	トピック分布
現代 IT 入門 A	0.041039962, 0, 0.17913537, 0.10145746, 0.672663, 0
グローバル社会動態発展演習 A	0, 0, 0, 0.9773268, 0, 0
国際関係論 A	0, 0.032327175, 0.44885013, 0, 0.5099458, 0

また、LDA トピックモデルでは自動的にトピック数を決められないので自身でトピック数を決める必要がある。その際の指標となるのが Perplexity と Coherence である。

3-2-1. Perplexity

Perplexity はモデルでの選択肢の数を表している。例えばある文書の 1 単語が隠されているとする。文書の語彙数が 10000 のときそこに入る単語の選択肢は 10000 である。LDA による Perplexity が 1000 のとき、それは LDA によって単語の選択肢の数を 1000 にまで減らしたことを意味する。このように Perplexity は小さいほどそのモデルの性能が良いことを示す。

3-2-2. Coherence

Coherence はトピック中の単語間類似度の平均値であり、トピックの質を表す。トピック全体の Coherence が高ければ良いモデルである。

3-3. コサイン類似度

コサイン類似度はベクトル空間において、2 本のベクトルがなす角度を表す指標である。以下の式(3)で求められる。1 に近ければ類似しており、0 に近ければ似てないことを表す。

$$\cos(\vec{q},\ \vec{d}) = \frac{\vec{q}\cdot\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\cdot\frac{\vec{d}}{|\vec{d}|} \qquad (3)$$

4. 提案手法

4-1.　シラバスデータの取得・整形

分析対象となるシラバスは神戸大学外部公開用シラバスのものを用いた。その中でも国際人間科学部の2016年から2020年の3501授業の中から科目名、時間割コード、開講年度、授業のテーマをBeautiful SoupとSeleniumを用いてスクレイピングした。授業のテーマのテキストデータにはJanomeを用いて形態素解析を行い、わかち書きをした。名詞が授業の特徴を表すと仮定し、わかち書きされたシラバスの単語群の中から名詞のみを抽出した。さらに、「それ、こと」などの授業の特徴を表さないと思われる単語や記号はストップワーズのリストを作り、それらを取り除いた。表1はその一部分である。

表2.　シラバスの一例

授業名	名詞
音楽文化史 1	エポックメーキング,音楽,作品,作曲,家,音楽,芸術,表現,様式,変遷,社会,文化,史,意味,考察
現代社会理論 A	貧困,共有,事態,人類,歴史,共同,性,基礎,近代,後,個人,化,過程,貧困,忘却,進展,現代,私,自己,認識,社会,帰結,私,キーワード,現代,時,空間,認識
情報リテラシー演習 1	オンライン,コミュニケーション,文書,処理,計算,基本,操作,方法,身,情報,機器,具体,活用,技能,習得

4-2.　トピックベクトル生成

トピックモデルの分析には、Pythonライブラリのgensimを用いた。LDAにおいてはトピック数は自動的には決まらず。事前に指定して行う必要がある。トピック数を決める際の指標として、PerplexityとCoherenceの2つを用いた。シラバスデータに対してトピック数を2~50に変化させ、PerplexityとCoherenceを求めプロットしたものが図1である。Perplexityは大きいほど良く、Coherenceは小さいほど良いとされるのでトピック数を6に設定した。トピック数6にしてLDAを実行し、トピックに属する単語をワードクラウドで表示したものが図2である。

図 1.　PerplexityとCoherenceのプロット結果

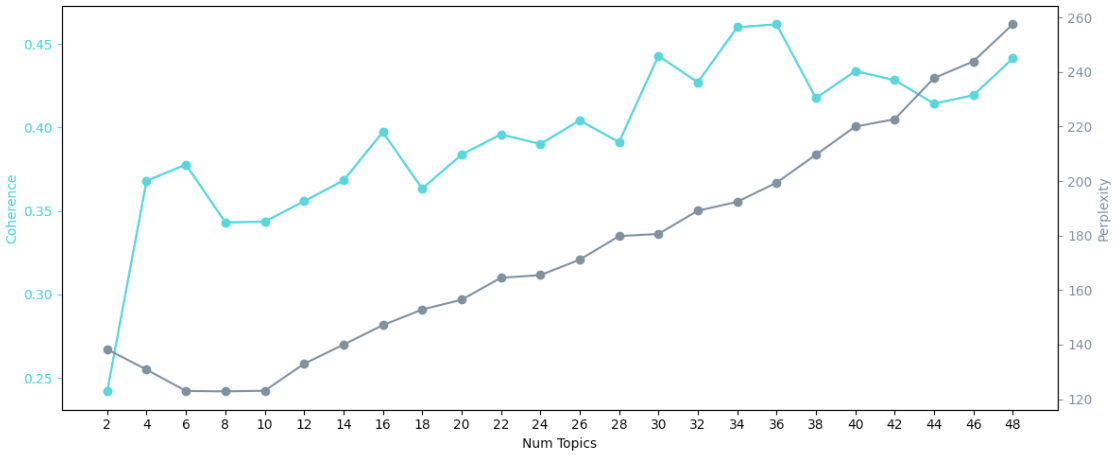


図2.　LDAによって生成された各トピックに属する単語をワードクラウドで表示したもの

生成したLDAモデルを用いて、各授業のトピックベクトルを求める。その後、各授業の科目名、開講年度、時間割コード、授業のテーマをわかち書きした名詞群、トピックベクトルの列を持つファイルを作成する。

4-3. レコメンド方法

4-3-1. 成績データ

提案するレコメンド方法では、先ほどのシラバスファイルと学生の成績データを用いる。学生の成績データは、うりぼーネットの履修成績照会にあるファイルを出力するボタンによって得られるCSVファイルを用いる。CSVファイルの成績の列は計算にできるように数値に変換する。GPAに合わせて、秀を4.2、優を4、良を3、可を2、不可を0に変換した。成績が合格の授業は基本的に必修授業であるため、今回はおすすめに関与しないように0とした。履修取り消しされた授業は成績データから除いた。

4-3-2. 嗜好性の取得

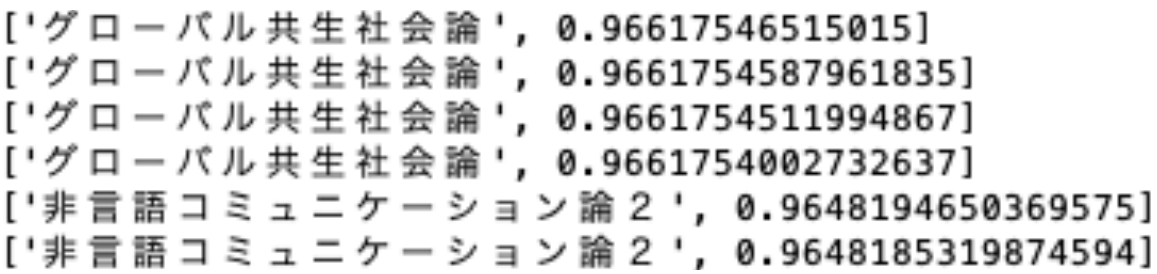
学生の成績データ内にある授業の年度と時間割コードによりシラバスファイルから検索し、該当する授業を探し出す。発見した授業のトピックベクトルそれぞれの値に成績データの値をかけ重みづけを行う。各授業のトピックベクトルに対してこの計算を行い、トピックごとに値を合計する。このようにして合計されたベクトルは学生の各トピックに対する嗜好性を表す。例えば、トピック1の値が大きければ、トピック1は得意だと考えられ、値が小さければ不得意であると考えられる。

4-3-3. レコメンド

おすすめしたい授業のトピックベクトルと学生の嗜好性ベクトルの類似度をコサイン類似度を用いて求める。コサイン類似度が大きい授業ほど嗜好性に合っているため、その学生

におすすめである。コサイン類似度が大きい授業から降順に並べ、上位の授業をおすすめ授業として学生にレコメンドする。実際にターミナル常に出力された授業の例が図3である。

図3.　ターミナルに出力されたレコメンドの一例（右の値はコサイン類似度）



5. 評価実験

5-1.　評価実験方法

レコメンド方法についての評価にあたっては、学生の履修履歴にある科目群の中からレコメンドし、レコメンドされた順番と成績の相関係数を見る方法を行った。成績とレコメンド順に正の相関関係があれば、レコメンド方法は学生の科目に対する得意・不得意に即したレコメンドができていると考えられる。実際のデータは、神戸大学国際人間科学部グローバル文化学部に所属する学生4名の履修成績データを用いた。

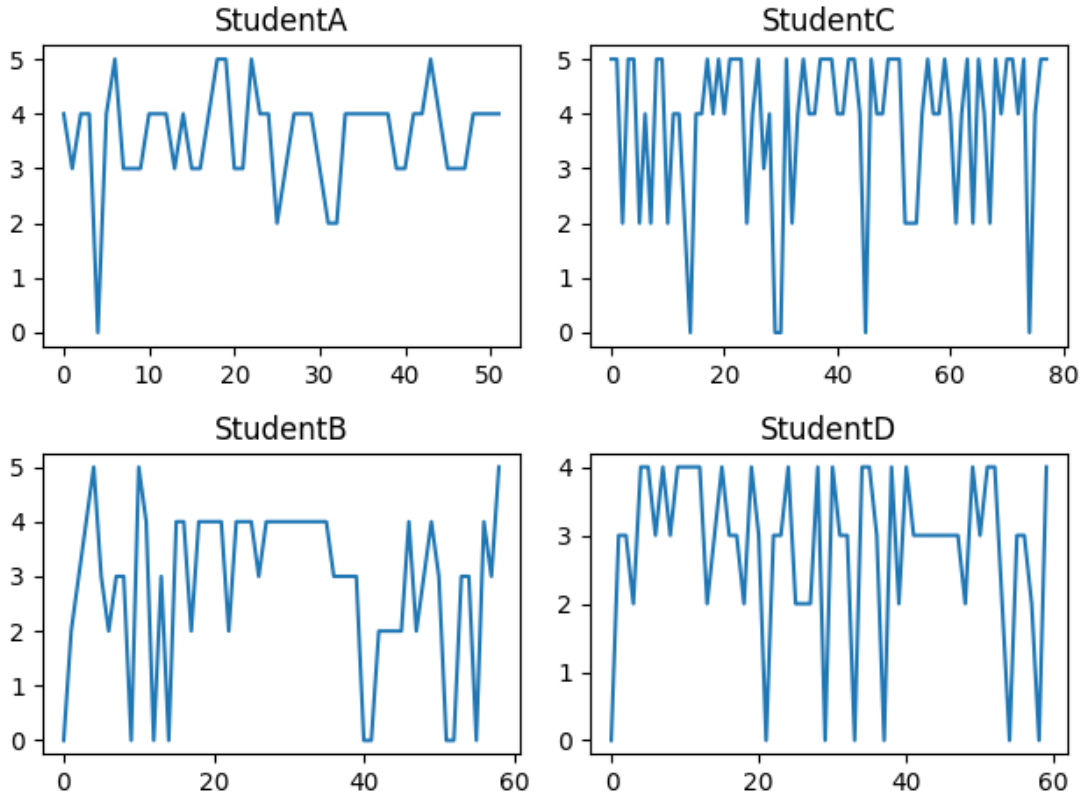
5-2.　実験結果

表2は相関係数を計算してみた結果である。最も相関係数が大きい学生の値でも約0.12であるためほとんど相関はないと考えられる。また、横軸をレコメンドの順位、縦軸を成績にして表にプロットしたものが図4である。おすすめ度が高い横軸の左部分に好成績の授業が、おすすめ度が低い右部分に成績が良くない授業が来るのが理想である。しかし図を見てわかるように、好成績の授業は左右に散らばっており、反対に成績の良くない授業も点在している。このことから、今回のレコメンド方法は学生の科目に対する得意不得意を反映できていないと言える。

表3.　成績とレコメンド順の相関係数

学生	相関係数
学生A	-0.0994319
学生B	0.08027
学生C	-0.0578176
学生D	0.116057

図4. プロット結果（横軸：レコメンド順, 縦軸：成績）



5-3. 考察

成績とレコメンド順に相関がない理由として大きく分けて3つの原因が考えられる。シラバスのテキストデータ、レコメンド方法、トピック以外の成績に影響する要因である。それぞれに対して考察していく。

5-3-1. シラバスデータ

今回使用したシラバスの授業テーマのテキストデータは授業によってその書き方や量がバラバラである。例えば近現代社会思想論Aの授業テーマは「近代社会をめぐる諸理論」のみであり、LDAモデル生成の際に用いられている単語は「近代、社会、理論」の3単語の

みである。この3単語から授業の特徴を捉えることは難しい。このようにシラバスの文章量が少なく、特徴を捉えられていない授業があるためおすすめの精度が低下していることが考えられる。

5-3-2. レコメンド方法

今回のレコメンド方法は授業のトピックベクトルに直接成績をかけて、トピックごとに合計して得られた嗜好性ベクトルと授業のトピックベクトルの角度が小さいものをレコメンドするという方法である。しかし、この方法以外にも学生が一番得意なトピックの値が大きい授業をおすすめする方法や他の学生の嗜好性ベクトルを用いて協調フィルタリング的におすすめする方法もなども考えられる。

5-3-3. トピック以外の成績に影響する要因

今回はトピックベクトルと成績のみを用いてレコメンドを行った。図5は得られた嗜好性ベクトルとレコメンドされた授業のトピックベクトルをターミナルに出力したものである。上に表示されてるほどレコメンド順が高い。これを見てみるとこの学生はトピック4の値が約46.6と最も大きくおすすめされている授業も全てトピック4の値が最も大きいことがわかる。今回のレコメンドシステムは学生が成績の良いトピックを学習し、それに適したレコメンドを行っている。図6は授業のトピックベクトルと成績を表示したものである。認知コミュニケーション論1, 2、近現代社会思想論A, Bはそれぞれトピックベクトルがほとんど相似している。しかし学生の成績は2と5、4と3など異なっている。このことから学生は授業のトピックベクトルが同じ授業でも成績にばらつきがあることがわかる。この要因としては、学生の履修したときの状況や成績評価基準の相違などが考えられる。例えば、体調の悪化や学外の用事などがあれば、授業への参加率は下がり成績が下がる。その他には、同じテーマの授業でも成績の評価がテスト形式からレポート形式に変われば成績が異なる場合が考えら

れる。このように学生の成績には授業のトピックのような内容以外も様々な要因が重なっているため、トピックを基準にしてレコメンドをすると成績を反映できないということが考えられる。

図5.　嗜好性ベクトルとおすすめされた授業のトピックベクトルの一例

```
嗜好性ベクトル
[24.146849085300005, 25.069839884999997, 30.411181599399995, 46.6232829211, 27.597299086, 30.0076713839]


おすすめの授業

コミュニティ創成論
[[0.0105243335, 0.0105209928, 0.010526286400000001, 0.9474309683000001, 0.0104595795, 0.010537861800000001]]
[[2019 '1H024' 4]]
---------------------------------
初年次セミナー
[[0, 0, 0, 0.9712190032, 0, 0]]
[[2018 '1H007' 0]]
---------------------------------
文化人類学２
[[0, 0, 0, 0.9765758514, 0, 0]]
[[2018 '2H310' 3]]
---------------------------------
情報リテラシー演習２
[[0, 0, 0, 0.9581028819, 0, 0]]
[[2018 '4H026' 4]]
---------------------------------
グローバルコミュニケーション発展演習Ａ
[[0, 0, 0, 0.9557955861, 0, 0]]
[[2019 '3H422' 4]]
---------------------------------
```

図6. 授業のトピックベクトルと成績

```
認知コミュニケーション論１
[[0, 0, 0, 0.23389753700000002, 0, 0.741119504]]
[[2018 '1H385' 2]]
---------------------------------
認知コミュニケーション論２
[[0, 0, 0, 0.2074543685, 0, 0.7588641644]]
[[2018 '2H385' 5]]

近現代社会思想論Ｂ
[[0.0420177877, 0.0419114567, 0.041926916700000004, 0.7902703285, 0.041764553600000004, 0.042108908300000004]]
[[2018 '4H312' 4]]
---------------------------------
近現代社会思想論Ａ
[[0.0420177579, 0.041911438100000004, 0.0419268943, 0.7902725935, 0.0417645462, 0.0421067774]]
[[2018 '3H312' 3]]
```

6. おわりに

6-1. まとめ

本研究では、LDA トピックモデルと成績を用いた履修レコメンドシステムを開発し、成績とレコメンド順の相関を求めることでその評価を行った。その結果、成績とレコメンド順には相関関係が見られず、今回のレコメンド方法では適切なレコメンドができていないことが確認された。

6-2. 今後の課題と展望

レコメンドがうまくいかなかった原因として、データ、レコメンド方法、トピック以外の要因の成績に対する影響を挙げた。データの問題は、一部の授業のデータの量が少なかったことに起因すると考えられる。そこで今回用いた授業テーマだけでなく、授業の概要と計画や成績の評価方法などを用いることによって、データの量と多様性を増やすことで解決できる可能性がある。レコメンド方法は、コサイン類似度以外にも、一番得意なトピックのみを考慮しておすすめする方法が考えられる。トピック以外の成績に影響する要因は、成績評価方法や学生の環境の変化などが考えられる。前者はデータの問題でも述べたようなシラバスの成績評価基準を用いることで考慮に入れることができる。後者は要因を特定することが難しい。履修成績にある授業に対して持っていたモチベーションをアンケートによって集め、モチベーションと成績の観点から分析することができるかもしれない。

トピックと成績の相関は今回の研究では見られなかったが、トピックの精度をあげることや多様なデータと組み合せることで、LDA トピックモデルと履修成績を用いた授業レコメンドができる可能性がある。また、今後は LDA トピックモデルに加えて。協調フィルタリングの技術を用いて、複数の学生の履修成績を考慮に入れた履修レコメンドシステムの研究を行っていきたい。

参考文献

[1] 竹森汰智, 亀井清華. “科目推薦のための doc2vec の応用方法の検討”. 情報処理学会. 2018

[2] 西森友省, 堀幸雄, 今井慈郎. “履修履歴を用いた科目推薦システム”. 情報処理学会. 2013.

[3] 奥村学, 佐藤一誠. “トピックモデルによる統計的潜在意味解析”. コロナ社. 2015.

LDA トピックモデルと履修成績を用いた履修レコメンドシステム

所属：国際文化学部

学籍番号：1686592C

氏名：宮崎仁弥

本論文では「LDA トピックモデルと履修成績を用いた履修レコメンドシステム」について記す。

現在学生が履修可能な授業数は多く、履修スケジュールを考えることは煩雑化した。そこで、履修スケジュールを考える時間の短縮と履修成績の分析による学生の得意・不得意なトピックの分析を目的として、本研究では LDA トピックモデルと履修成績を用いた科目レコメンドシステムを構築、その評価を行った。