ARG WI2 No.11, 2017
WI2-2017-22

関連科目を検索するための シラバスデータ分析手法の検討

吉崎 辰悟*， 越智 洋司**

*近畿大学大学院 ， **近畿大学

ochi@kindai.ac.jp

概要 学生が自分の必要とする講義や参考書を探し出す手段として，シラバスを利用する方法が挙げられる．シラバスとは，大学で開講される授業や講義の大まかな内容を示したものであり，講義の目的や各時限の授業内容などが記載されている．しかし現状では，大量のシラバスデータから必要な情報を探し出す事は難しく，科目ごとの類似性や関連性を把握することは困難である．そこで，シラバスデータから関連科目を導き出し，関連科目データベースを作成することで学生が講義の内容をより深く理解するための手助けになるのではないかと考えた．本稿では，近畿大学のシラバスデータを対象に，文章ベクトルから関連科目を導出する手法について述べる．

キーワード シラバスデータ，関連科目，Doc2Vec

1 はじめに

近年，多くの大学で「シラバス」が導入されており，学生はシラバスを利用することによって自分の必要とする講義や教科書の検索を行う．シラバスには講義の大まかな内容や授業目標等が示されており，この記載内容に基づいて講義が実施される．しかし現状では，大量のシラバスデータから必要な情報を探し出す事は難しく，科目ごとの類似性や関連性を把握することは困難である．そこで，シラバス情報の「授業内容」の項目に着目し，この項目が類似する科目を分析することで関連科目が導出できるのではないかと考えた．本稿では，近畿大学のシラバスデータを対象に，文章ベクトルから関連科目を導出する手法について述べる．

2 関連研究

これまで，シラバスに関する様々な研究が行われてきた．川場ら[1]は，ウェブシラバスを公開している高等教育機関のシラバスを収集・分析し，汎用型ウェブシラバスシステムを開発している．また，堀ら[2]はシラバスのテキストデータから TF-IDF により特徴抽出を行うことで，履修決定支援や個人のカリキュラムの特徴を見出す研究を行っている．

文書から特徴量を抽出する方法として単語の出現頻度を用いる TF-IDF が挙げられる．しかし，これは単語の語順を考慮できないため，関連する文書が必ずしも似たベクトルになるとは限らないという問題があった．そこで本稿では，文書内部の単語を比較しベクトルを割り当てることで類似度の導出を行うDoc2Vecを用いて，シラバスデータの分析をする．

3 関連科目の導出手法

本研究では，シラバス本文中の単語を形態素解析し，その類似性から関連科目を導き出す．なお，類似性の分析についてはDoc2Vecを使用する．以下，実装の方法について述べる．

(1) 形態素解析

形態素解析とは，文章をある単語に区切り，辞書を利用して品詞や内容を判別することである．本研究では，形態素解析器として「kuromoji」を使用した．シラバスの文章から名詞のみを抽出し，メタデータの作成を行う．

(2) Doc2Vec

Mikolov ら[3]によって提案された Doc2Vec は，単語をベクトル化して定量化することにより文書の類似度やベクトル計算などを実現する．通常，1文書ごとに単語を羅列して文書間を比較するのに用いられるが，本研究では1科目ごとの単語集合を1文書分として科目間の類似性を抽出する．なお，本研究では実装ライブラリとして chainer[4]を利用した．

4 ユーザインタフェース

UI の作成として「GWT Bootstrap」を用いる．GWT Bootstrap はレスポンシブル Web デザインに対応しており，ブラウザの横幅サイズを判断基準としてレイアウトデザインを柔軟に調節する事が可能である．本システムでは検索画面，科目一覧画面，関連科目検索結果画

面を作成．科目一覧画面で科目を指定すると，Doc2Vec により導出した関連科目が表示される．

5 評価実験

本研究では，シラバス内の文書が類似する科目を分析することで関連科目が導出できるものであると仮定し，Doc2Vec により導出した関連科目（以下，類似度関連科目）の妥当性を検証する．評価方法は，シラバスに明記されている関連科目（以下，シラバス関連科目）と類似度関連科目が対応しているかを確認する手法[5]で行う．なお，本検証では近畿大学理工学部電気電子工学科で開講されている 189 科目分のデータを対象とする．評価には式(1)の適合率と式(2)の再現率を用いる．式中の変数は表1の混合行列に対応している．

$$\text{適合率} = \frac{TP}{TP + FP} \quad \text{式(1)}$$

$$\text{再現率} = \frac{TP}{TP + FN} \quad \text{式(2)}$$

表 1 混合行列

		類似度関連科目	
		あり	なし
シラバス関連科目	あり	TP	FN
	なし	FP	TN

本システムでは，学習した結果の類似度が高い順に科目名が表示される．この際，類似度のどの値までが関連科目として妥当であるのかを定める必要がある．類似度関連科目について，類似度の設定値を 0.3 以上~0.7 以上で変化させた時の適合率と再現率の推移を図 1 に示す．

類似度が 0.3 以上の科目を類似度関連科目と定めると，より多くの科目が類似度関連科目の対象となるため，TP の値が大きくなり再現率が高くなる．一方で類似度を 0.7 以上とした場合，TP の値が減少するため適合率・再現率は低下するが，関連性の強い科目のみを類似度関連科目として抽出することが可能である．

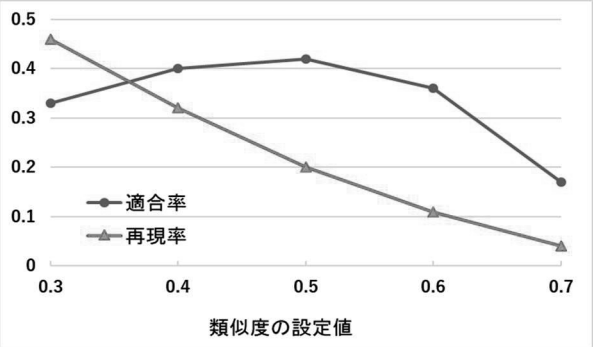

図1 類似度関連科目の精度推移

表 2 Epoch 数を変更した時の学習結果

Epoch	適合率	再現率	F 値
5	0.382	0.378	0.380
10	0.368	0.329	0.347
15	0.421	0.371	0.395
20	0.411	0.375	0.392
50	0.369	0.318	0.341
100	0.379	0.339	0.358
200	0.358	0.330	0.344
400	0.326	0.292	0.308

また，Doc2Vec では次元数や Epoch 数などのパラメータを任意の値に設定することが可能である．表 2 に，Epoch 数のみを 5~400 までの値に変更して学習させた時の適合率と再現率，その F 値を示す．

Epoch 数が 15~20 の時に適合率が 0.4 以上となり，F 値も高くなる．しかし Epoch 数を上げすぎると，適合率・再現率共に低くなる傾向がある．これは，データに対して過学習が発生しているためと考えられる．

6 おわりに

本稿では，関連科目を検索するためのシラバスデータの分析手法と，関連科目導出方法について述べた．類似度関連科目として導出した科目の中で，シラバスに関連科目として記載されていればその関連度は自明であるが，「類似度関連科目には存在するが，シラバスには関連科目として記載されていない科目(FP 科目)」の妥当性については本評価実験では判断していない．シラバス関連科目は教員の主観が入ったものであり，関連性が高いにも関わらずシラバスに関連科目として記載されていない科目も存在すると考えられる．今後，FP 科目に関する教員の主観評価を行う予定である．

参考文献

[1] 川場隆，土屋健，小柳恵一：汎用型ウェブシラバスシステムの開発，日本教育工学会論文誌，35(Suppl.)，pp61-64, 2011

[2] 堀幸雄，中山堯，今井慈郎：カリキュラムの特徴抽出と時間割の要約生成，情報知識会誌，Vol.20, No.2, pp201-206, 2010

[3] Tomas Mikolov, Quoc Le: Distributed Representations of Sentences and Documents Proc, Of ICML2014, pp1188-1196, 2014

[4] Chainer: A flexible framework for neural networks: http://chainer.org/, accessed Nov.14, 2016

[5] 佐々木健太郎，土田正明，田村晃裕：期待再現率における期待適合率最大化モデルの学習方法，第 21 回言語処理学会，pp481-484, 2015