Vol.2015-CE-132 No.16
Vol.2015-CLE-17 No.16
2015/12/5

日本十進分類を用いた カリキュラム比較のための講義自動分類

増田 勝也[1,a)]

概要：本研究ではカリキュラムの比較に利用することを目的として，シラバステキストを用いて講義を分野の観点で自動的に分類することを目的とする．講義の分類先として図書の分類である日本十進分類 (NDC) を利用し，シラバス情報を用いて講義を一定の分類体系に分類することで，大学間でのシラバスやカリキュラムの比較を行うことが容易となる．また分類に NDC を利用することで講義と図書を容易に結びつけることができ，大学における図書館の利用を促すことができる．本研究では人手で付与された図書に対する NDC および書籍タイトルなどの書誌情報を学習データとして利用し, 機械学習手法を用いることで講義シラバスに対し自動的に NDC の分類を付与する．精度評価は図書の分類に対して行い，講義の分類結果は実例により示す．

1. はじめに

本稿では日本十進分類 (Nippon Deciaml Classification, NDC) を分類体系として講義シラバスを分野の観点から自動的に分類する手法を提案する．大学教育においては「チューニング」と呼ばれる，大学間で学習内容や到達目標・コンピテンス等を共有し，学生や研究者の大学間交流を促進しようとする試みが近年議論されている．そのためには大学間でカリキュラムや講義を比較し，それぞれの大学の講義間の関連性やカリキュラム間の相補関係などを明示的に示すことが重要であると考えられる．そこで本研究ではカリキュラムや講義の比較にむけて，講義シラバスのテキスト情報を用いて講義を分野の観点から標準的な体系を用いて分類することを目的とする．それらの分類に利用することで講義間・カリキュラム間の関連性，対応関係を明らかにすることが可能となる．標準的な分野体系としては図書の分類体系である日本十進分類 (NDC) を用いる．カリキュラムの標準的な分類としては，特定の分野に対しては情報学分野における J07[1] のようなカリキュラム標準が存在するが，全学問分野に対するカリキュラムの標準的な分類は存在しないため，学問分野に対する総合的な分類体系として NDC を用いる．NDC は主に日本で使用されている図書分類であるが，国際的に使用されている図書の分類であるデューイ十進分類 (Dewey Decimal Classification, DDC) と一定の対応を取ることができるため [2]，同様の方法で国際的にもカリキュラムの比較を行うことが可能となる．また図書の分類である NDC を用いて講義を分類することで講義と図書館の連携を行うことができ，学生の図書館の利用促進に繋がると考えられる．分類手法としては機械学習手法の一つである Random Forest[3] を用い，図書の書誌および目次データから分類器を作成する．作成した分類器を図書の書誌データに適用し精度を計るとともにシラバステキストに対して適用し実際のシラバスがどのように分類されるかを提示する．

[1] 東京大学
The University of Tokyo
[a)] masuda@he.u-tokyo.ac.jp

2. 関連研究

シラバスデータを利用した講義の自動分類，カリキュラムの比較のための可視化等は様々な手法で研究が行われている．由谷はシラバスから専門用語を抽出しその重みを用いて科目間の類似度を計算することで科目間の関連性の分析を行っている [4]．井田や野澤は講義をシラバスを用いてクラスタリングし，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より，カリキュラム間の比較を行っている [5], [6]．太田は東京大学工学部のシラバスを対象としてシラバスから抽出された専門用語を利用し機械学習手法によりシラバスの分類を行っている [7], [8], [9]．関屋はカリキュラムの比較を目的として，LDA と Isomap を用いて講義を二次元平面上に射影し可視化を行っている [10]．ここでは情報系分野を対象とし，標準的なカリキュラムを基準として可視化を行うことで，大学のカリキュラム間の比較を可能にしている．

また一方で NDC をある対象に自動付与する研究は様々な対象についておこなわれている．NDC は本来図書の分類であるので，図書の自動分類の研究が多数行われている．石田は図書の検索を目的として，単語単位での NDC に対する重み付けを利用したベクトルの類似度により分類を行っている [11]．また後には目次データや帯のデータを利用し，機械学習手法 (Support Vector Machine) を用いて自動分類を行っている [12]．また畑田はニューラルネットワークを用いて自動分類を行っている [13]．宮田は NDC の階層構造を考慮に入れ，機械学習手法を用いて NDC 一桁ずつの分類を行っている [14].

図書以外では，図書館のレファレンスデータベースに対する NDC の自動付与が行われている．レファレンスデータベースは図書館員の資料・情報の探索のためのデータベースであり，国会図書館においてレファレンス協同データベースとして NDC とともにデータが蓄積されている．原田は国会図書館のレファレンス協同データベースを対象として，三種類の機械学習手法を用いて自動分類を行っている [15]．また荒井は単語を NDC に対する重みを要素とするベクトルに変換し，機械学習手法を用いて自動分類を行っている [16]．また図書館とは関連のない事柄の NDC による自動分類としては人物の分類が行われている [17], [18], [19]．これらは Web での人物のカテゴリ検索を目的として，人物に関する Web ページのテキストを用いて NDC の分類を行っている．

また講義に対し NDC 分類を付与する試みは明治大学において行われている [20]．この試みでは講義と図書の連携を目的として授業シラバス約 1,200 件に NDC を人手で付与し，OPAC 検索での利用を可能にしている．ここでは，講義と図書を関連付けることで学生の図書館利用を促すことを目的としており，本研究とは目的は異なるが NDC と講義を結びつけるという観点では同じである．明治大学において人手により行われた NDC の付与を本研究は機械学習手法を用いて自動的に行う．

3. 日本十進分類を用いた機械学習による自動分類

3.1 日本十進分類 (NDC)

日本十進分類は日本において広く使用されている図書分類法である．国際的に使用されているデューイ十進分類法を基に，日本に関連した項目を重視するなどして作成された分類法である．分類記号としてはアラビア数字を用い，各分類では 10 区分に分類しながら大分類から小分類へと階層的に順次細分していく．分類記号は三桁目までの 1000 分類を基本とし，それ以上は必要に応じて分類を行う．本稿においては基本分類である三桁目の分類を対象として分類を行う．

3.2 機械学習による自動分類

提案する手法では機械学習を用いて分類を行うために，分類対象となるテキストを特徴ベクトルに変換し，そのベクトルを対象として機械学習による分類器の作成および分類を行う．テキストから特徴ベクトルへの変換方法としては [16] で提案されている手法を改変した手法を用いる．まずテキストに対し形態素解析 MeCab[21] を用いて形態素解析を行ない，その中から「用語」として名詞連続を抽出する．各用語 t に対し，以下の要素からなるベクトル $v_t = (w_t^{c_1}, ..., w_t^{c_n})$ を求める．

$$w_t^{c_i} = \frac{tf_t^{c_i}}{\sum_{c_j} tf_t^{c_j}}$$

ただし，c_i はある NDC 分類，$tf_t^{c_i}$ は用語 t の NDC が c_i である書籍内での全ての出現頻度である．すなわちこのベクトルは用語 t がどの NDC の書籍によく出現するかを表現したベクトルといえる．

学習データ内の全用語 t について上記のベクトルを求めたうえで，分類対象となるテキストの特徴ベクトルを以下の方法で作成する．まず対象となるテキストからテキスト中の用語集合 S を用語のベクトルを求めた時と同様の方法で抽出する．そしてテキスト d の特徴ベクトル v_d は以下の要素からなるベクトル $v_d = (w_d^{c_1}, ..., w_d^{c_n})$ とする．

$$w_d^{c_i} = \sum_{t \in S} w_t^{c_i}$$

すなわち対象テキスト中の用語の特徴ベクトルの和をそのテキストの特徴ベクトルとする．また最後に正規化を行い，テキストの長さに依存しない特徴ベクトルとする．本手法の利点は，特徴ベクトルの要素数が限定されることである．単純に用語の頻度を特徴ベクトルの要素として利用する手法では，用語の種類が多くなるにつれ特徴ベクトルの要素数が多くなり，また低頻度語が多いためベクトルが疎になりがちであり，分類器の学習が困難になってしまう．本手法のように用語を抽象化することでベクトルが疎になることを防ぎ，データが増えても安定した学習を行うことが可能となる．

上記の方法で作成したテキストの特徴ベクトルを用いて，機械学習を用いて分類器を作成する．本研究では機械学習手法として Random Forest[3] を用い，実装としては Weka[22] を使用した．Random Forest は多数の決定木を用いた集団学習アルゴリズムであり，分類時には決定木集合の多数決により最終的な分類を行う．

3.3 階層的分類

本稿においては，NDC の基本分類である上位三桁を対象として分類を行うが，一度に 1000 分類の中から 1 分類を推定するのは困難であるため，NDC の階層構造を利用して一桁 (10 分類) ごとの分類を行なった．すなわち，ま

Vol.2015-CE-132 No.16
Vol.2015-CLE-17 No.16
2015/12/5

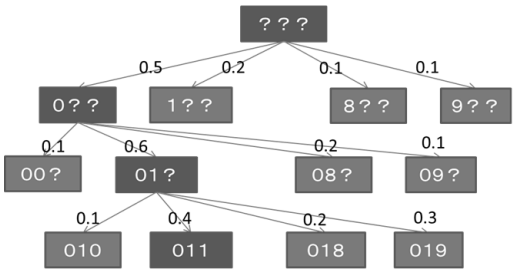

図 1 分類手法の模式図

ずは一桁目を分類し，その分類内でさらに二桁目を分類，さらにその分類内で三桁目を分類することで三桁全てを分類する手法とした．分類の模式図を**図 1** に示す．

手法の詳細は以下のとおりである．機械学習手法を用いているため，手法は大きく学習フェーズと分類フェーズに分けることができる．学習フェーズでは，各桁の分類を行う分類器を前章で述べた手法により作成する．まず一桁目については学習データに対し一桁目の値を対象として分類を行う分類器 C_{*--} を作成する．次に二桁目については，一桁目の分類 n_1 により 10 分割したデータそれぞれに対し，その中で二桁目の値を対象として分類を行う分類器 C_{n_1*-} を作成する．すなわち 10 区分に分類を行う分類器を，一桁目の値ごとに合計 10 個作成する．また三桁目の分類も同様に一桁目・二桁目の値 n_1,n_2 により 100 分割したデータに対し三桁目の値をラベルとして分類を行う 100 個の分類器 $C_{n_1n_2*}$ を作成する．以上により 10 区分の分類を行う分類器が計 111 個作成され，以下の分類フェーズにおいて順次適用していくことで三桁の NDC 一つに分類を行うことができる．

分類フェーズでは，分類対象のテキストに対し前章の方法で特徴ベクトルを作成し上の桁から順番に分類を行っていく．まず一桁目を分類器 C_{*--} を用いて推定する．次に二桁目を，推定された一桁目の分類 m_1 に対応する分類器 C_{m_1*-} を用いて分類する．三桁目も同様に一桁目・二桁目の推定された値 m_1,m_2 に対応する分類器 $C_{m_1m_2*}$ により分類する．これにより三桁の NDC 分類を行うことができる．図 1 の例では，まず NDC が不明な入力テキストに対し一桁目の分類器 C_{*--} を適用し，最も確率値が高い一桁目として「0」(確率値 0.5) を得る．次に二桁目の分類には分類器 C_{0*-} を使用し一桁目が「0」である NDC の中で二桁目の分類を推定する．図中では二桁目は「1」が最も確率値が高く，二桁目の分類として「1」を得る．続いて三桁目の分類には分類器 C_{01*} を使用し二桁目までが「01」である NDC の中で分類を行ない，三桁目の分類として「1」を得，最終的に入力テキストの NDC として「011」が出力される．

また，各桁の分類において一分類のみを出力するのではなく，分類器から出力された確率値の高い方から指定された個数の分類を利用して次の桁の分類を行う．その際次の桁の分類においては前の桁での確率値を重みとして利用し，その桁での分類による確率値と掛けあわせて最終的な分類の確率値として確率値の高い順に指定された個数の分類を出力する．これにより複数の NDC を出力することが可能となり，また分類の探索範囲の面においても推定する NDC の範囲を広げ，上位の桁で推定を誤ったことで全体として推定を誤るということを防ぐことができると考えられるが，上位の桁での出力分類数を増やすことで下位の桁での分類器の適用個数が増えるため，分類時間がその分長くなる．なお今回は機械学習手法として Random Forest を，特徴ベクトルとして NDC に対する重みベクトルを利用したが，他の分類手法を使用することも可能である．

表 1 分類精度

	候補数 1	候補数 3	候補数 5	候補数 10
一桁	0.420	0.420	0.420	0.420
二桁	0.226	0.106	0.102	0.102
三桁	0.118	0.055	0.054	0.054

4. 実験

4.1 実験データ

今回の実験に使用する書誌情報データとしては，学習データとして国立国会図書館のデジタル化が行われた書籍の書誌・目次データ 848,829 件，評価データとして東京大学駒場図書館の直近 3 年分の OPAC データ 20,913 件を利用した．国会図書館データについては書籍のタイトルおよび目次データを分類に使用する対象テキストとし，駒場図書館データについては目次データが存在しないため，書籍のタイトルのみを分類対象のテキストとした．また，実際のシラバスへの適用例として，東京大学授業カタログ[*1]にて公開されているシラバスデータを利用し，国会図書館データで学習された分類器を用いて分類を行なった．分類に使用する項目は講義の内容が記述されていると考えられる「授業の目標概要」「授業計画」「授業科目名」の三項目を使用した．

4.2 実験結果

上記の分類手法による分類結果を**表 1** に示す．精度は各桁において最も確率値の高い分類と正解との適合割合として計算する．三桁の精度がおよそ 10%であるが，精度が低い原因として，学習データの不足等が考えられる．特に学習データ中の語彙不足ならびに NDC の分布の偏りが原因であると考えられる．また各桁での候補数を増やすと精度が低下していることがわかる．候補数を増やすことで NDC の探索範囲が広がったゆえに，正しい NDC を推定していたところが誤った NDC の推定になってしまっている．

また実際にシラバスの自動分類を東京大学授業カタログ

[*1] http://catalog.he.u-tokyo.ac.jp/

表 2 講義の NDC 分類例

講義名	自動付与された NDC
憲法第１部	323(憲法)
金融論	338(金融．銀行．信託)
哲学概論	331(経済学．経済思想)
宗教学概論	375(教育課程．学習指導．教科別教育)
考古学概論 (1)	335(企業．経営)
解剖学	491(基礎医学)
数学１A	375(教育課程．学習指導．教科別教育)
物理数学	410(数学)
コンピュータ科学	509(工業．工業経済)
建築構造計画概論	511(土木力学．建設材料)
都市建築史概論	332(経済史・事情．経済体制)
遺伝子科学	491(基礎医学)
薬学概論	375(教育課程．学習指導．教科別教育)

で公開されているシラバスと対象として行なった．シラバスの分類の際には各桁での候補数を 1 として分類を行なった．分類結果の例を**表 2** に示す．分類された分野が適合している講義と，適合していない講義とでかなりのばらつきがある．また大学の講義のシラバスであるため 375(教育課程．学習指導．教科別教育) と分類されている講義がかなりの数存在していた．これは講義の説明に一般的に使用される用語に強く影響されていると考えられる．シラバスへ自動分類を適用し，すでに稼働している東京大学授業カタログでの可視化システムにおいてデータとして利用することで分類に基づく可視化・クラスタリングを行うことが可能となる．

5. 考察

本稿での実験において精度が低い原因としては，以下の点が考えられる．

学習データの偏り

実験で使用したデータは学習データが国会図書館の書誌・目次データ，テストデータが駒場図書館の書誌データであるが，データにおける NDC の分布に偏りがあることが低精度の原因の一つと考えられる．**表 3** に学習データにおける NDC 一桁目の分布を示す．表から分かる通り，NDC の一桁目においてもデータ数にかなりのばらつきがある．また国会図書館のデータは年代が古いものが多く，最新の書籍のデータは含まれていないため，直近 3 年分の駒場図書館のデータとは語彙の分布が異なっている．また NDC の分布についても，例えば 007(情報科学) のように国会図書館データには含まれないが駒場図書館データには含まれている，といった分布の異なりが存在している．また，特に 9 類 (文学) の書籍が 3 類 (社会科学) に次いで多く，文学作品においては多様な語彙が使用されているため

表 3 NDC 分類別書籍数

NDC(類)	文献数 (国会)	文献数 (駒場)
0(総記)	27,017	751
1(哲学)	74,508	2,517
2(歴史)	94,790	2,223
3(社会科学)	204,542	6,912
4(自然科学)	65,902	2,479
5(技術)	69,066	1,132
6(産業)	87,816	585
7(芸術)	57,675	1,249
8(言語)	24,984	935
9(文学)	142,534	2,130

表 4 NDC 分類別書籍数 (9 類)

NDC(網)	文献数 (国会)	文献数 (駒場)
90(文学)	5,104	181
91(日本文学)	111,085	1,145
92(中国文学)	4,625	103
93(英米文学)	8,727	314
94(ドイツ文学)	3,403	160
95(フランス文学)	5,655	124
96(スペイン文学)	135	20
97(イタリア文学)	321	14
98(ロシア文学)	3,414	48
99(その他)	315	21

表 5 NDC 分類別書籍数 (91 網)

NDC(目)	文献数 (国会)	文献数 (駒場)
910(日本文学)	4,883	378
911(詩歌)	29,024	196
912(戯曲)	3,902	12
913(小説．物語)	49,609	267
914(評論．エッセイ)	6,355	114
915(日記．書簡)	3,785	26
916(記録．手記)	106	43
917(箴言．寸言)	72	3
918(作品集)	7,317	97
919(漢詩文)	5,782	9

その点が分類に悪影響を与えているとも考えられる．

階層的分類手法の問題点

各桁での候補数を増やすことにより探索範囲を広げ，精度が向上できると考えていたが，実験結果では候補数を増やすことにより逆に精度が低下している．これは確率値の計算と学習データの偏りに一因があると考えられる．提案手法では，各段での分類の確率値を掛けあわせて最終的な確率値としているが，NDC の分布の偏りのため，この方法では逆に正しくない結果となる場合がある．例えば 913(小説．物語) は 9(文学)→91(日本文学)→913(小説．物

語) という階層になっているが，**表 4** から分かるように 9 類においては 91(日本文学) の書籍数が他に比べ非常に多くなっている．そのため，提案手法の階層的分類において 9 類の中で二桁目の分類を行う際には，91(日本文学) の確率が非常に高くなる．また同様に 91(日本文学) 内での三桁目の分類においても，**表 5** から分かるように 911(詩歌) および 913(小説，物語) の書籍数が他に比べて非常に多いため，これらの分類の確率が非常に高くなる．そのため，各桁での確率値の積で全体の確率値とした場合，他の類に比べ 913(小説．物語) の確率値が相対的に高くなってしまう．各桁での候補数を増やす際には，このような分類間の分布の偏りを考慮する必要があると考えられる．

6. おわりに

本論文ではカリキュラムの比較を目的として，講義シラバスに対する自動分類システムを提案した．分類体系として図書の分類に利用される日本十進分類を用い，機械学習手法として Random Forest を用いて，テキスト中の用語の頻度を基に自動分類を行なった．

今後の課題としてはまずはデータの整備があげられる．考察で示したとおり学習データについては現在の使用方法を含め解決すべき点があると考えられるので，それを解決するためのデータの整理や新規データの取得が考えられる．また本来の目的としている対象はシラバスデータの分類であるためその分類の評価のために実際のシラバスを用いた正解データ，すなわちシラバスの人手による NDC 分類が必要であると考えれる．現時点ではシラバスデータに NDC が付与された正解データは得られていないため，人手により作成された実際のシラバスデータでの精度を評価し，その上でさらなる精度の向上を目指したい．

参考文献

[1] 情報処理学会情報処理教育委員会：情報専門学科におけるカリキュラム標準 J07, https://www.ipsj.or.jp/12kyoiku/J07/J0720090407.html.
[2] 高木貞治：図書館における書誌分類 : DDC と NDC 間の分類対応表の作成 : 総合目録データベースを利用して，大学図書館研究， Vol. 57, pp. 31–38 (1999).
[3] Breiman, L.: Random Forests, *Mach. Learn.*, Vol. 45, No. 1, pp. 5–32 (online), DOI: 10.1023/A:1010933404324 (2001).
[4] 由谷真之，森 幹彦，喜多 一：N-007 電子シラバスを用いた大学教養教育のカリキュラム分析 (N 分野:教育・人文科学)，情報科学技術フォーラム一般講演論文集， Vol. 4, No. 4, pp. 315–316 (2005).
[5] 井田正明，野澤孝之，芳鐘冬樹：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).
[6] 野澤孝之，井田正明，芳鐘冬樹，宮崎和光，喜多 一：シラバスの文書クラスタリングに基づくカリキュラム分析システムの構築，情報処理学会論文誌， Vol. 46, No. 1, pp. 289–300 (2005).
[7] 太田 晋，美馬秀樹：課題志向別シラバス自動分類システムの設計と実装，自然言語処理， Vol. 16, No. 4, pp. 91–106 (2009).
[8] 太田 晋，美馬秀樹：10-106 課題志向別シラバス自動分類システムの開発，工学・工業教育研究講演会講演論文集， Vol. 21, pp. 172–173 (2009).
[9] 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).
[10] 関谷貴之，松田源立，山口和紀：LDA と Isomap を用いた計算機科学関連カリキュラムの分析，情報処理学会論文誌， Vol. 54, No. 1, pp. 423–434 (2013).
[11] 石田栄美：図書を NDC カテゴリに分類する試み, *Library and information science*, Vol. 39, pp. 31–45 (1998).
[12] 石田栄美，宮田洋輔，神門典子，上田修一：目次と帯を用いた図書の自動分類，情報処理学会研究報告. FI,. 情報学基礎研究会報告， Vol. 82, pp. 85–92 (2006).
[13] 畑田 稔：ニューラルネットワークによる図書の自動分類，全国大会講演論文集，Vol. 57, pp. 360–361 (1998).
[14] 宮田洋輔，石田栄美，神門典子，上田修一：NDC の階層構造を利用した図書の自動分類の試み，2006 年度日本図書館情報学会春季研究集会発表要綱，pp. 51–54 (2006).
[15] 原田隆史，江藤正己，大西美奈子：レファレンスデータに対する NDC の自動付与，情報知識学会誌， Vol. 17, No. 2, pp. 61–64 (2007).
[16] 荒井俊介，辻 慶太：機械学習を用いたレファレンスデータへの NDC の自動付与, 情報知識学会誌，Vol. 25, No. 1, pp. 23–40 (2015).
[17] 浦 芳伸，村上晴美：NDC を用いた人物ディレクトリの開発，全国大会講演論文集，Vol. 2011, No. 1, pp. 651–653 (2011).
[18] Murakami, H. and Ura, Y.: People search using NDC classification system, *Proceedings of the fourth workshop on Exploiting Semantic Annotations in Information Retrieval, ESAIR 2011, Glasgow, United Kingdom, October 28, 2011*, pp. 13–14 (2011).
[19] Murakami, H., Ura, Y. and Kataoka, Y.: Assigning Library Classification Numbers to People on the Web, *Information Retrieval Technology - 9th Asia Information Retrieval Societies Conference, AIRS 2013, Singapore, December 9-11, 2013. Proceedings*, pp. 464–475 (2013).
[20] 中林雅士：NDC 分類を使った授業と図書館資料の連携，大学図書館研究， Vol. 95, pp. 64–74 (2012).
[21] 工藤 拓，山本 薫，松本裕治：Conditional Random Fields を用いた日本語形態素解析，情報処理学会研究報告自然言語処理 (NL)，Vol. 2004, No. 47, pp. 89–96 (2004).
[22] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H.: The WEKA Data Mining Software: An Update, *SIGKDD Explor. Newsl.*, Vol. 11, No. 1, pp. 10–18 (online), DOI: 10.1145/1656274.1656278 (2009).