様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月　 ４ 日現在

機関番号：１２６０１
研究種目：基盤研究(B)（一般）
研究期間：2017～2019
課題番号：１７Ｈ０１８３７

研究課題名（和文）自動収集した大量のシラバス情報を用いたカリキュラムの定量的分析とその応用

研究課題名（英文）Quantitative Analysis and Application of Curriculum Using a Large Amount of Automatically Collected Syllabus Information

研究代表者

関谷　貴之（Sekiya, Takayuki）

東京大学・情報基盤センター・助教

研究者番号：７０３２３５０８

交付決定額（研究期間全体）：（直接経費）　　7,900,000円

研究成果の概要（和文）：本研究は、カリキュラムの特徴や傾向を、主観的要素を排した形で把握することを目標とする。まずシラバスの提供形態を次の3つに分類した：シラバス情報を記載したウェブページへのリンク集となる Link Type、複数のシラバス情報がまとまった Whole Type、そしてシラバスを集約して提供する Database Type である。その上で、各タイプに応じた判定モデルによるツール、Google Search API 経由で求めたシラバスに強く関わるページから、それにリンクされたページを取得するクローラ、判定ツールの分析結果などを保持するデータベースで構成されるシラバス収集支援システムを開発した。

研究成果の学術的意義や社会的意義
高等教育機関のカリキュラムを客観的な基準で分析するためには、その内容を端的に表すシラバスを大量に集めて分析することが望ましい。しかし、その分析のためには多大な労力を要する。本研究では、シラバスの提供形態の分類と、それに合わせたシラバスの判定から得られた知見に基づく、シラバス収集支援システムを開発したことで、今後大量のシラバスを効率的に集める一助となる。

研究成果の概要（英文）：This study aims to understand the characteristics and trends of the curriculum without any subjective elements through a quantitative analysis. In order to realize the goal, we obtained a large amount of syllabus information, and categorized them into the following three types: Link Type, which is a collection of links to syllabus web pages; Whole Type, which is a collection of multiple syllabus information; and Database Type, which gathers and provides syllabus information offered by educational institution. Then, we developed a syllabus collection support system which consists of a decision tool using a decision model by the SVM for each classification, a syllabus crawler that searches for pages that are strongly related to the syllabus based on keywords via the Google Search API and combines it with a complementary generic crawler, and a database which holds the web pages and their meta-information, and analysis results by the decision tool.

研究分野：教育支援システム

キーワード：シラバス　ウェブクローラ　機械学習　カリキュラム

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

カリキュラムの設計や改良には、対象とする当該学問分野で教育すべき項目や教育の結果得られる学習効果を十分に検討する必要がある。しかし、このような分析や検討には、当該分野の専門的知識と教育工学的知識が要求され、容易な作業ではない。従って、カリキュラム分析手法を開発することは、カリキュラムの設計や改善の一助となる。

我々は Latent Dirichlet Allocation (LDA) や Isomap などの機械学習の技術を応用して、情報処理学会が提案する計算機科学における学部生向けのカリキュラム標準 J07-CS 等を基準として用い、シラバスからカリキュラムの全体構造を把握するマップを作成する方法を研究した。また、ACM と IEEE が提案する Computer Science Curricula 2013 (CS2013) を基準として、米国の主要大学のカリキュラムを分析した。

２．研究の目的

教育機関のカリキュラムを定量的に分析することで、カリキュラムの特徴や傾向を、主観的要素を排した形で把握することを一つの目的とする。そのために、これまで我々が取得してきたシラバス情報を訓練データとするウェブクローラを開発して、半自動的にシラバスをより大量に取得することを試みる。

３．研究の方法

計算機科学分野のカリキュラムを調査の対象とするに当って、国際的な大学のランキングの一つである Times Higher Education (THE) WORLD UNIVERSITY RANKINGS (https://www.timeshighereducation.com/world-university-rankings) 2018 with computer science as subject (以下、「THE2018CS」という。)として公開されている大学の一覧を用いた。大学院生の協力で、それらの各大学で計算機科学のシラバス情報を公開しているウェブサイトを探索し、人手によるデータ抽出の正解例とした。

並行して、大学のウェブサイトのドメイン名を指定することで、大量のウェブページをクローリングして、シラバス情報を公開するウェブページを抽出するために、既存の検索エンジンや汎用ウェブクローラ、Support Vector Machine (SVM) の判定モデルを用いたツールなどを組み合わせたシステムを構築した。その上でシステムを評価した。

４．研究成果

(1) THE2018CS 各大学の CS 関連のウェブサイトの URL 及び収集範囲を設定した上で、汎用的なウェブクローラ Scrapy (https://scrapy.org) でクローリングしたウェブページを用いることで 2020 年 3 月現在 301 大学から合計約 570 万ページ以上を取得した。また、予備実験として、CS 分野を専攻する大学院生 2 名に、取得したウェブページからシラバス情報を抽出することを依頼した。抽出作業に当たって、過去の研究で取得済みのシラバスが記載されたウェブページを正解データとして、ウェブページ内のテキストに基づく Support Vector Machine (SVM) の判定モデルを構築して、クロールしたウェブページの判定結果を示している。THE2018CS の上位 100 校をそれぞれ 50 校ずつ前述の大学院生に割り当てた。
その結果、合計で約 60 校からシラバス情報を抽出することができた。

(2) (1)で抽出したウェブページを分析したところ、その内容やリンク構造からシラバス情報を提供するウェブページ(ウェブサイト)は以下の 3 種類に分類できることが分かった。

- **Link Type: 学部や学科等で開講される科目ごとのシラバスページへのリンク集となっているページである。 例えば図 1 左は Carnegie Mellon University の CS undergraduate courses を記載した Link Type のページ**

(http://coursecatalog.web.cmu.edu/schoolofcomputerscience/undergraduatecomputerscience/) である。

- Whole Type: 複数のシラバス情報が比較的短くまとまっているページである。例えば図 1 中は University of California, Irvine の Department of CS のシラバス情報をまとめた Whole Type のページ (http://catalogue.uci.edu/donaldbrenschoolofinformationandcomputersciences/departmentofcomputerscience/#courseinventory) である。
- Database Type: 学部や学科、大学全体で開講される科目のシラバス情報を検索できるデータベースの入り口となるページである。例えば図 1 右は、University of Melbourne のシラバス情報などを検索できるサイト Handbook のトップページ (https://handbook.unimelb.edu.au/) である。

なお、それぞれの Type に当てはまるページを有する大学の数の比率は、大まかに Link : Whole : Database = 3 : 2 : 1 であった。

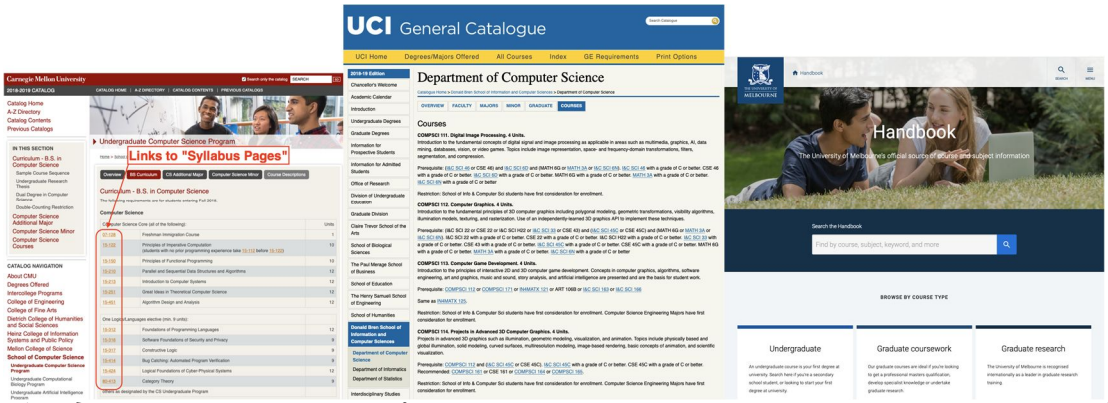

図 1 左から Link Type のサンプル (Carnegie Mellon University)、Whole Type のサンプル (University of California, Irvine)、及び Database Type のサンプル (University of Melbourne)

(3) Link Type / Whole Type / Database Type を判定するための SVM のモデルを構築した。Link Type 判定用のモデル “TR-LINK” の作成手順と評価結果について特に詳しく述べる。大学院生が見つけたページの中で、クローラでも取得できたページ 22 個をモデルの正解データとして用いた。また、過去にクローリングしたウェブページについて、シラバスとは関係ないことを目視で判定したページ 500 個を不正解データとして用いる。ウェブページを処理して判定モデルを作成するに当っては、HTML タグを除去したテキストの抽出に “html2text” を、stop word の除去や bag of words への変換、そして SVM 等に “scikit-learn” (https://scikit-learn.org/stable/)を用いている。なお、TR-LINK の作成に当たっては、Link Type のページが多数のリンクを含んでいることを考慮して、ウェブページ内のアンカータグに含まれる URL を、ページから抽出したテキストに追加している。TR-LINK の性能評価に当たって、leave-one-out cross validation (LOOCV) を実施した結果を表 1 に示す。
今回用いた scikit-learn の SVM 用の関数である LinearSVC のパラメータ C は 10^{-3} から 10^{3} の範囲で、0.1 の時に f1-score が最も高くなり、このとき不正解データをすべて判定できた一方、正解データのうち 18 ページを判定できた。

表 1: TR-LINK の性能 (LOOCV)

precision	recall	f1-score	support
1.000	0.818	0.900	22

TR-LINK の評価に当たって、leave-one-out cross validation (LOOCV) を実施した結果を表 1 に示す。
LinearSVC のパラメータ C は 10^{-3} から 10^{3} の範囲で、0.1 の時に f1-score が最も高くなり、このとき不正解データをすべて判定できた一方、正解データのうち 18 ページを判定できた。加えて、TR-LINK を用いてシラバスページを見つけることができるかを検討するべく、正解データからある大学のページのみを除外して作成した SVM のモデル “TR-LINK-self” で、その大学からクローリングしたウェブページを判定することを試みた。

もし、除外した正解データのウェブページを Link Type のページとして適切に判定できたとすれば、 confidence score で並べ替えた当該大学のウェブページの中で上位に現れ、その順位を当該大学から取得したウェブページの総数で割ると小さな値となることが期待される。22 大学すべてについて行った結果の平均値は 13.7% であった。これらの結果から、今回作成した SVM の判定モデル TR-LINK は、Link Type のウェブページを探すためにある程度役に立つといえるだろう。

(4) シラバスの収集を支援するシステムを開発した。図 2 にシステムの概念図を示す。本システムは、クローラ、判定ツール、データベース、の 3 つの部分からなる。クローラについては、大学のトップページからリンクされたウェブページを広く取得した上で後述の判定ツールでシラバスに関連するページを抽出する方法なども検討したが、クローリングに要する作業時間などを考慮した結果、キーワードに基づいてシラバスに関連する少数の入口となるウェブページを、対象となる大学のドメイン以下に限定して、 Google Search API を用いて抽出した上で、その入口からリンクされた大量のウェブページを取得する方法を採用することとした。そこで、先の Google Search API と 汎用ウェブクローラ Scrapy (https://scrapy.org) とを組み合わせたクローラを開発した。判定ツールは、シラバス情報の提供形態に合わせた Support Vector Machine の判定モデルを選択して、クローラが取得した大量のウェブページの中からシラバスとの関わりが強いページを見付ける。データベースは、ウェブページのコンテンツ、クローラによるページ取得に関する条件や取得日時、及び判定ツールの判定結果などのメタ情報を保持する。コンテンツの保持用の mongoDB (https://www.mongodb.org) と、その他の各種情報を保持する PostgreSQL (https://www.postgresql.org) で実装した。過去に大学院生がシラバス情報を抽出したように、ある分野の知識を備えた研究者と同様に、指定した分野のシラバス情報を抽出できるかについては、引き続き評価中である。

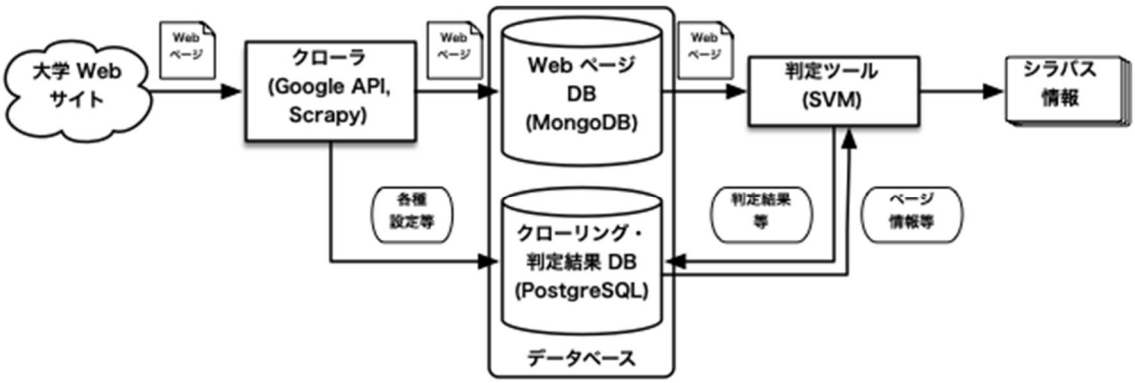

図 2 シラバス収集支援システム

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Matsuda Yoshitatsu、Sekiya Takayuki、Yamaguchi Kazunori	４．巻 10638
２．論文標題 Discovery of Interconnection Among Knowledge Areas of Standard Computer Science Curricula by a Data Science Approach	５．発行年 2017年
３．雑誌名 Neural Information Processing	６．最初と最後の頁 186-195
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1007/978-3-319-70139-4_19	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Sekiya Takayuki、Matsuda Yoshitatsu、Yamaguchi Kazunori	４．巻 0
２．論文標題 A web-based curriculum engineering tool for investigating syllabi in topic space of standard computer science curricula	５．発行年 2017年
３．雑誌名 2017 IEEE Frontiers in Education Conference (FIE), Indianapolis, IN, USA	６．最初と最後の頁 1-9
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/FIE.2017.8190598	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Matsuda Yoshitatsu、Sekiya Takayuki、Yamaguchi Kazunori	４．巻 26
２．論文標題 Curriculum Analysis of Computer Science Departments by Simplified, Supervised LDA	５．発行年 2018年
３．雑誌名 Journal of Information Processing	６．最初と最後の頁 497～508
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.2197/ipsjjip.26.497	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Matsuda Yoshitatsu、Yamaguchi Kazunori	４．巻 29
２．論文標題 A Unifying Objective Function of Independent Component Analysis for Ordering Sources by Non-Gaussianity	５．発行年 2018年
３．雑誌名 IEEE Transactions on Neural Networks and Learning Systems	６．最初と最後の頁 5630～5642
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TNNLS.2018.2806959	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Sekiya Takayuki、Matsuda Yoshitatsu、Yamaguchi Kazunori	4．巻 1
2．論文標題 Investigation on University Websites for Semi-automated Syllabus Crawling	5．発行年 2019年
3．雑誌名 IEEE Frontiers in Education Conference (FIE)	6．最初と最後の頁 1-7
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/FIE43999.2019.9028479	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Dai Jinan、Yamaguchi Kazunori	4．巻 1
2．論文標題 Compact and Robust Models for Japanese-English Character-level Machine Translation	5．発行年 2019年
3．雑誌名 Proceedings of the 6th Workshop on Asian Translation	6．最初と最後の頁 36-44
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/D19-5202	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 松田源立
2．発表標題 適応的独立成分分析によるノイズ除去と特徴抽出
3．学会等名 第20回情報論的学習理論ワークショップ，2017.11.8～11，東京大学 本郷キャンパス
4．発表年 2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	松田　源立 (Matsuda Yoshitatsu) (40433700)	成蹊大学・理工学部・准教授 (32629)	

６．研究組織（つづき）

	氏名 （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	山口　和紀 (Yamaguchi Kazunori) (80158097)	東京大学・大学院総合文化研究科・教授 (12601)	