3N2-2

科目区分ダイアグラム検索システムにおけるテキスト類似度に基づく科目推薦機構の試作
Implementing a Function for Recommending Academic Subjects based on Textual Similarity for a Curriculum Module Search System

宮脇克典[*1] 池田雄斗[*1] 福本加奈恵[*1] 水野創太[*1] 白松俊[*1]
Yoshinori Miyawaki Yuto Ikeda Kanae Fukumoto Sota Mizuno Shun Shiramatsu

[*1] 名古屋工業大学 大学院工学研究科 情報工学専攻
Department of Computer Science, Graduate School of Engineering Nagoya Institute of Technology

In 2016, “Creative Engineering Education Program” was newly established in Nagoya Institute of Technology. In this program, students select diversified academic subjects from thirteen faculties and design their curriculum by themselves. However, it is difficult to select academic subjects adequately for the new students because they have no deep specialized knowledge. We implement a system for searching curriculum modules for students’ self-design of curriculum and a function for recommending suitable lectures based on the descriptions of students’ targets and academic subjects’ syllabuses. In the recommendation function, we calculate similarities on the basis of two methods: TF-IDF and Paragraph Vector. Our experimental result shows that the similarities based on TF-IDF is more suitable than that based on Paragraph Vector.

1. はじめに

名古屋工業大学では，幅広い工学知識を持ち多角的に物事を観察し社会に役立つ人材の育成，輩出を目的とした新学科「創造工学教育課程」が設立された．これは各々の掲げる目標に応じて，自らのカリキュラム分野横断的に設計する教育課程である．カリキュラム設計において選択できる部門には電気・機械工学，生命・応用化学，物理工学，情報工学，そして社会工学がある．学生は各自，設定した目標について学習目標を記述する．これをCプランと言い，それを元に幅広い専門分野から学生自身が受講科目を選択し，カリキュラムを設計する．カリキュラムは1年次に設計し，適宜再設計と更新を行う．補助に冊子のダイアグラムが配布されているが，初学者が目的の科目を探すことは困難が予想される．そこで本研究では，WEB 上でダイアグラムが閲覧できるようなシステムを開発し，学生の補助を図る．

システムの要件として，1)学科・区分ごとのダイアグラム表示機能，2)科目のキーワード検索機能，3)シラバスと連携した閲覧機能，4)学生の目標記述に沿った科目推薦機能，が挙げられる．また，科目推薦機能について，入学初年度の学生にはCプランに対して履修すべき科目を適切に選択することは難解である．そのため，TF-IDF と sentence2vec の 2 つの手法により C プランとシラバスの類似度を求め，科目推薦を行い，どちらがより適切に科目推薦を行えるか，比較・実験を行う．

2. 科目区分ダイアグラム検索システムの実装

本システムのシステム構成を図 1 に示す．クライアント側でダイアグラムの科目や区分が選択されることで，サーバ側にリクエストが送信される．そのリクエストから SPARQL クエリを受け取った RDF ストア(Stardog)がデータを JSON 形式で返信する．そこで WEB API がシステムで運用しやすい JSON-LD 形式に変換をする．このデータによりダイアグラムを描画する．また，図 2 がシステムの実際の画面となっている．

連絡先：宮脇克典，〒466-8555 名古屋市昭和区御器所町 名古屋工業大学つくり領域白松研究室,yoshinori@srmtlab.org

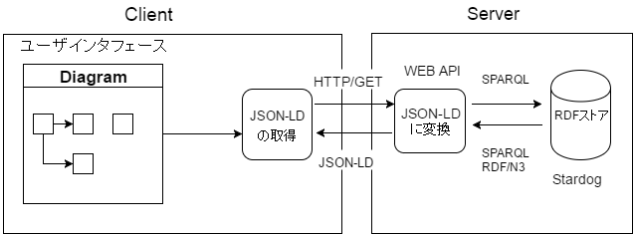

図 1: システムの概要

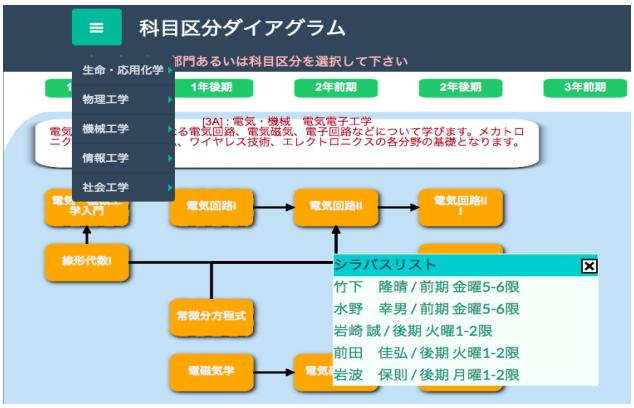

図 2:システムの使用画面

2.1 科目データの形式

ダイアグラムを描画するためには科目データが必要となる．本システムは，シラバス公開システムや履修システムとの連携を考慮すると，相互運用性が求められる．また，特定の科目は，前提科目という，受講する際に前もって修めておくべき科目が存在し，ダイアグラムではそれらの関係を矢印で結ぶ．したがって，本システムでは Linked Data を用いることが適切であると判断した．表 1 は科目クラスのプロパティである．科目の記述には，Linked Science プロジェクトの Teaching Core Vocabulary[LinkedScience 2012]を用いた．接頭辞 cplan で始まるプロパティは新たに追加したものである．

表 1: 科目クラスのプロパティ

プロパティ	説明
teach:hasTitle	講義名
teach:bookingNumber	講義番号
teach:module	科目の属する区分番号
teach:academicTerm	開講時期
cplan:department	学科番号
cplan:row	描画に用いる行番号
cplan:lineStyle	描画時の線の種類

2.2 サーバ側のシステム構成

サーバ側には，クライアントと RDF ストア間でデータをやり取りするための WEB API，そして RDF ストアの Stardog がある．

2.3 WEB API

WEB API の機能について述べる．クライアントからリクエストを受け取ると SPARQL クエリを生成し，Stardog に問い合わせる．Stardog から返信されたデータを JSON-LD 形式に変換する．クライアントから受け取るリクエストには(1)学科(2)区分(3)シラバス公開システムへのリンク(4)キーワード検索の 4 つがある．

2.4 Stardog

Stardog とは Complexible 社が提供しているセマンティックグラフデータベースであり，本システムでは 250 万トリプルが利用可能な無償版を使用している．

2.5 クライアント側のシステム構成

クライアント側の実際の画面は図 2 である．機能としては，(1)学科・区分ごとにダイアグラムを表示する機能，(2)科目のキーワード検索機能，(3)科目のシラバスリンクがある．

3. 学生の目標記述を用いた科目推薦機能

学生の学習目標記述による科目推薦について述べる．創造工学教育課程の学生は C プランという学習目標を記述する．これと科目シラバスについて類似度を算出し，類似度の高い科目を推薦する．

この類似度を求める際に，(1)TF-IDF と(2)sentence2vec の 2 つの手法を用いる．C プランは PDF 形式のため，Google ドライブの OCR 機能を使用し，不足や不適切な場合は改めて記述した．

3.1 TF-IDF による類似度算出

MeCab を用いて C プランとシラバスの名詞を抽出し，TF-IDF を求め，そこから C プランとシラバスのコサイン類似度を算出する．類似度に閾値を設定し，上位の科目を推薦する．

3.2 sentence2vec による類似度算出

sentence2vec は段落ごとに 1 文書として解釈を行い，また単語は半角スペースで区切られている必要がある．そのための前処理として MeCab を用いて分かち書きしている．その後，sentence2vec を用いて C プランとの類似度の高い科目を推薦する．sentence2vec は demo.py という Python ファイルを実行することでベクトル化を行う．

本研究のパラメータ設定について述べる．C プラン 42 文書，科目シラバス 1062 文書をまとめた文書をコーパスとする．コーパスを解析する Word2Vec ではベクトルの次元数を表す size を

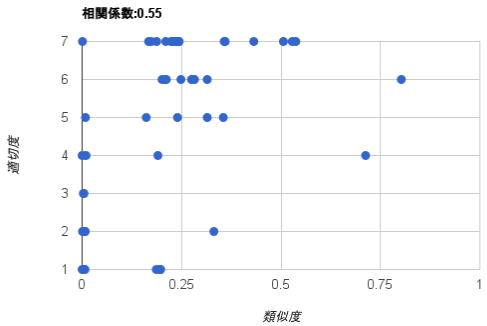

図 3: TF-IDF による推薦評価

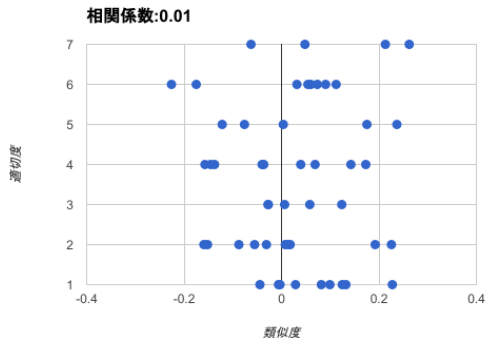

図 4: sentence2vec による推薦評価

100 に，ベクトルの隠れ層の非線形ノードの計算アルゴリズムについて Skip-gram と C-BOW(Continuous Bag-of-Words)の選択を行う sg を 0 に，単語を出現回数によって足切りを行う min_count は 5 に設定した．つまり，コーパスの単語ベクトルの次元数は 100，隠れ層の非線形ノードの計算アルゴリズムは C-BOW，出現回数が 5 回未満の単語は足切りを行うというせっていである．

sentence2vec とは Paragraph Vector[Quoc 2014]を Kang Longbiao が Python により実装したプログラムであり，これを用いて単語や文章の類似度を算出することができる．Paragraph Vector とは Facebook の Mikolov らが word2vec[Mikolov 2013]を発展させた，段落をベクトル表現する手法出ある．類似度は sentence2vec のメソッドを使用する．アルゴリズムは TF-IDF と同様にコサイン類似度を使用している．

4. 評価・考察

4.1 システム仕様ログの考察

科目区分ダイアグラム検索システムを 2016 年 11 月に創造工学教育課程の学生向けに公開した．利用統計によると，キーワード検索機能の利用は 16 件と最も少なく，学科表示機能が 52 件，区分表示機能 270 件，そしてシラバス閲覧機能が 700 件と最も多い結果であった．このことから，学生は学科・区分についての学習分野について推測できる一方で，個々の科目がどのような授業内容であるか，学習目標に寄与できる授業内容であるかということについてわからないため，シラバスの閲覧件数が多いことが推察される．また，キーワード検索機能の利用が少ない点について，特定分野の専門知識が十分身についておらず，目的の科目を検索するための専門用語がわからないためであると推察される．

4.2 科目推薦の実験・比較

Cプランとシラバスの類似度をTF-IDFとsentence2vecの2つの手法によって求め，それぞれで類似度上位の7科目，下位からランダムに3科目の計10科目を選択し，Cプランに対して推薦されたシラバスの内容が適切であるかを，1(不適切である)～7(適切である)の7段階で評価実験を実施した．5種類のCプランに対して8件のアンケートを行った．

TF-IDFによる推薦結果は図3，sentence2vecによる推薦結果についてテキストのノイズを取り除き改めて推薦評価を行った結果である．テキストのノイズを取り除く前に行った評価実験において，選択された数字が5以上であるとき正しいとして適合率が最も高くなる閾値を求めた.TF-IDFは閾値を0.34に設定した場合に適合率が93.6%となった．
sentence2vecは閾値を0.34に設定した場合に適合率が35.2%となった．また,7段階評価と類似度の相関係数を求めたところ，TF-IDFは+0.55となり中程度の正の相関が見られた一方で，sentence2vecは+0.07とほぼ相関が見られなかった．これは，シラバスでは教師名と開講日，Cプランでは学生や引用文献の著者などのノイズ除去を適切に行えていなかったことや，使用したCプランやシラバスの文章量が少なく，単語の足切りに引っかかった重要な単語が多かった事が，精度を悪化させてしまった原因である考えられる．ただし，TF-IDFによる推薦では，Cプランの内容が「感性工学」という知能情報分野について学ぶという目標に対して，「感性と社会」という，アメリカでの黒人社会の歴史について学ぶ，知能情報とは遠い内容の科目が推薦されることや，科目名のみで授業内容などの詳細が一切記述されていない科目が推薦されることがあった．これはその単語のTF-IDFが高くなってしまったためだと考えられる．

また，テキストから人名などのノイズを除去し，sentence2vecについての類似度と評価について改めて算出した結果が図4となっている．相関係数は+0.01となり，ノイズ除去前と同様，ほぼ相関が見られない結果になった．

5. おわりに

本研究では創造工学教育課程の学生のカリキュラム設計補助の為に，1)学科・区分ダイアグラムの表示機能，2)科目のキーワード検索機能，3)シラバス公開システムと連携した閲覧機能，を持つ科目・区分ダイアグラム検索システムの実装，並びに4)創造工学教育課程の学生の学習目標記述による科目推薦の検討を行った．本システムでは4)の科目推薦機能は未実装となっている．今後はダイアグラム検索システムに科目推薦機能の実装を課題とする．また，推薦機能について，Cプランとシラバスのノイズ処理を厳密に行い，改めてより精度の高い実験を行う．今回の評価実験では，知識が身についていない学部1年を想定し，Cプランに推薦されたシラバスが適切かどうかという質問を行ったため，基礎的な知識が身につき，カリキュラムの再検討を行う学生に対しての推薦は想定していない．そのため，推薦された科目が目標に対して役立つかということについても実験・評価を行いたい．

謝辞

本研究の一部は科研費若手(B)(25870321)の支援を受けた．

参考文献

[LinkedScience 2012] Teaching Core Vocabulary Specification, http://linkedscience.org/teach/ns/

[Quoc 2014] Quoc V. Le, et al. Distributed Representations of Sentences and Documents, CoRR, abs/1405. 4053, pp. 1 - 9, 2014.

[Mikolov 2013] Mikolov, Tomas, et al. Efficient estimation of word representations in vector space,arXiv preprint arXiv:1301.3781, 2013a, pp1-12.