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　　元 年　　６ 月 ２１ 日現在

機関番号：１２１０２
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００４３８

研究課題名（和文）Wikipedia閲覧者に対する図書推薦

研究課題名（英文）Book recommender system for Wikipedia readers

研究代表者
辻　慶太（Tsuji, Keita）

筑波大学・図書館情報メディア系・准教授

研究者番号：３０３３３５４５

交付決定額（研究期間全体）：（直接経費）　　1,200,000円

研究成果の概要（和文）：Wikipediaの閲覧者に対して，閲覧中の項目に関連した図書を推薦するシステムを構築した。システムの利用者としては主に，ラーニング・コモンズや大学図書館において，Wikipediaで学習や調べ物をしている学生を想定している。学生は本システムが推薦する図書を館内で閲覧することで学習内容を深めることができる。
まずWikipediaの各項目のNDCカテゴリーを決定する機械学習手法を開発し，次に推定されたNDCカテゴリーと同じカテゴリーの図書から推薦図書を決定する手法を開発した。学生被験者に推薦図書を評価してもらったところ概ね好評であった。

研究成果の学術的意義や社会的意義
Wikipediaの閲覧者に対して，閲覧中の項目に関連した図書を推薦するシステムを構築する。システムの利用者としては主に，ラーニング・コモンズや大学図書館において，Wikipediaで学習や調べ物をしている学生を想定する。学生は本システムが推薦する図書を館内で閲覧することで学習内容を深めることができる。本研究では対象Webページは Wikipediaに限定するが，将来的にはWebページ全体を対象とし，各ページに関連する図書を推薦するようにしたい。本研究にはWebと図書館資料を結びつける意義がある。

研究成果の概要（英文）：Book recommender system for Wikipedia readers was developed. System users are assumed to be university students who are studying in university libraries and learning commons. Students can learn deeply by reading books recommended by the present system.
Machine learning method to determine NDC (Nippon Decimal Classification) category for each Wikipedia article was developed. Then, method to determine which books should be recommended from those whose NDC categories are the same as Wikipedia artiecles'. Subject students evaluated favorably the recommended books and thus the effectiveness of the present system was proved.

研究分野：図書館情報学

キーワード：図書推薦

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

## １．研究開始当初の背景

現在，Wikipedia は調べ物や学習における重要な情報源となっている。大学図書館において学生は，サーチエンジンの検索結果から Wikipedia のページを選ぶ場合が多いこと，Wikipedia の項目からリンクが張られているページを閲覧する場合が多いことも示されている。もし Wikipedia の各項目で，その項目に関連する図書館資料，特に図書が表示されたら，学生はその図書を閲覧し学習を深めることができる。ブラウザでそのような図書推薦を実現するアドオンは今のところ存在しない。また図書を推薦する手法の研究は多く成されてきたが，Wikipedia や閲覧中の Web ページに基づく手法などは提案されていない

研究代表者はこれまで機械学習によって図書推薦を行う手法の研究に取り組んできた。ただそこでは学生が自身の興味関心を，図書を 1 冊選んで表明することを想定してきた。即ち，OPAC で学生が図書を 1 冊クリックしたら，その図書に関連する図書が推薦できることを目指してきた。だがそのような機能は，(1)学生が OPAC に飛ぶこと，(2)図書を 1 冊選ぶこと，を前提とする。これは先述のような，学生はまずサーチエンジンに飛び，Wikipedia を見るという現実に反する。いわば学生に，通常よりも負担を強いる想定となっている。そこで本研究では，Wikipedia を閲覧している学生に，その項目に関連する図書を機械学習によって推薦する手法を開発したい。

## ２．研究の目的

Wikipedia の閲覧者に対して，閲覧中の項目に関連した図書を推薦するシステムを構築する。システムの利用者としては主に，ラーニング・コモンズや大学図書館において，Wikipedia で学習や調べ物をしている学生を想定する。学生は本システムが推薦する図書を館内で閲覧することで学習内容を深めることができる。

## ３．研究の方法

### 3.1　Wikipedia ページの入手と名詞の分散表現の獲得

(a-1) 2017 年 8 月 1 日の Wikipedia のダンプをダウンロードした。ファイルサイズは約 14GB であった。

(a-2) 上記ファイルから ns タグが 0 ではない，即ち“ Main/Article ”ではないページ及びリダイレクションのページを除去した。残ったページは 1,070,202 であった。

(a-3) 2017 年 8 月 10 日現在の mecab-ipadic-NEologd dictionary を辞書とした Mecab ver. 0.996 で形態素解析を行った。上記ファイル中には異なりで 5,512,620 個の名詞があった。

(a-4) 上記名詞に関して gensim の Word2vec を用いて 200 次元の分散表現を獲得した。

### 3.2　各 Wikipedia ページへの NDC の付与

(b-1) 先ほどの(a-2)で得たページから，ISBN を明記しながら「参考文献」を挙げているページを抽出した。そのようなページは 50,375 あり，ISBN は異なりで 45,151，延べで 117,219 であった。

(b-2) 上記 ISBN を国立国会図書館の OpenSearch に入力し，各 ISBN に関して完全な書誌データ 40,647 個を入手した。

(b-3) 各ページに関して引用図書の NDC を特定した。
(b-4) 最も多い NDC が 3 個以上で，かつそれが 8 割を超える場合，その NDC をそのページに付与されるべき NDC とみなした。そのようにして 5,385 個のページと NDC のペアが得られた。
(b-6) 5,385 個のペアは 4,835 個の訓練データと 500 個のテストデータに分けた。

CNN への入力は，各ページの(1)タイトル，(2)カテゴリ，(3)本文，の中の名詞とした。(1)(2)については先頭の 5 つの名詞とし，(3)については最も TF-IDF 値が高い名詞 10 個とした。先ほどの(a-4)で得られたこれら名詞の分散表現（200 次元）と，0 から成る同様に 200 次元のベクトルを行とする 50×200 の行列を構成し，CNN への入力とした。

CNN は Tensorflow ver. 1.0 で構築した。畳み込みニューラルネットワークのパラメータとしては，tf.nn.conv2d の strides は[1, 1, 1, 1]とし，padding は VALID とした。tf.nn.conv2d の後には tf.nn.relu と tf.nn.max_pool を用いた。tf.nn.max_pool のパラメータ ksize，strides，padding はそれぞれ，[1, 入力チャネルの高さ－フィルターの高さ＋1, 1, 1]，[1, 1, 1, 1]，VALID とした。出力層には softmax 関数を用いた。ミニバッチ，エポック数，dropout rate，learning rate はそれぞれ 100，300，0.5，0.0001 とした。SVM には LIBSVM ver. 3.23 を用いた。カーネルには RBF カーネルを用い，easy.py で最適なパラメータ C，γを決定した。上記の方法で各 Wikipedia ページに対して NDC を付与した。

### 3.3　推薦図書の決定

T 大学が所蔵する 621,129 から，各ページに対して推薦図書を決定した。推薦図書の候補としては 2006 年から 2017 年にかけて少なくとも年 1 回は借りられている図書を用いた。そして各 Wikipedia ページに対して，そのページの推定 NDC と同じ NDC を持つ図書を推薦図書候補とした。さらにそこから推薦図書 3 冊ずつを決定するために再び CNN と SVM を用いた。決定した推薦図書は学生被験者に評価してもらった。

## ４．研究成果

まず Wikipedia の各項目の参考文献に挙げられている図書の NDC（日本十進分類法）カテゴリーからその項目の NDC カテゴリーを決定し，それを学習データとして参考文献がない項目の NDC カテゴリーを推定できる手法を開発した。具体的には畳み込みニューラルネットワークで推定するための最適なパラメータを特定し，精度は 90%近くに達した。

次に推定された NDC カテゴリーと同じカテゴリーの図書の中から，図書タイトルと項目タイトル・カテゴリー・本文との類似度に基づいて，推薦すべき図書を決定する手法を開発した。ここでは畳み込みニューラルネットワークよりも SVM の方が有効であることが分かった。学生被験者に推薦図書を評価してもらったところ概ね好評であった。

## ５．主な発表論文等

〔雑誌論文〕（計　0　件）

〔学会発表〕（計　2　件）

(1) Tsuji, Keita (2017) "Automatic Classification of Wikipedia Articles by Using

Convolutional Neural Network," Proceedings of the 9th Qualitative and Quantitative Methods in Libraries International Conference (QQML 2017). (25th May, 2017 at The Savoy Hotel, Limerick, Ireland), 8p. (No Pagination).

(2) Tsuji, Keita (2016) "Books Cited in Wikipedia: Possibility to Use their Nippon Decimal Classification Categories for Book Recommendation," Proceedings of the 7th International Conference on E-Service and Knowledge Management (ESKM 2016). (12th July, 2016 at Kumamoto City International Center, Kumamoto, Japan) p.1196-1197.

**〔図書〕（計　0　件）**

**〔産業財産権〕**
**○出願状況（計　0　件）**

**名称：**
**発明者：**
**権利者：**
**種類：**
**番号：**
**出願年：**
**国内外の別：**

**○取得状況（計　0　件）**

**名称：**
**発明者：**
**権利者：**
**種類：**
**番号：**
**取得年：**
**国内外の別：**

**〔その他〕**
**ホームページ等**

## ６．研究組織