

小学校での地域学習 NIE に適した Web ニュース記事推薦システム

関 伸也[†] 安藤 一秋[‡]

香川大学大学院工学研究科[†] 香川大学創造工学部[‡]

1. はじめに

NIE (Newspaper in Education) は、新聞を教材として活用する取り組みのことであり、小学校や中学校を中心に幅広い教育機関で実施されている。各地域の NIE 推進協議会が発行している小学校での実践報告[1]には、地域学習の一環として地元や周辺地域に関する記事を用いて NIE を実践する例が多いことや、NIE を継続的に行うことで、児童の読解力・語彙力や社会への興味関心が向上することなどが述べられている。

小学校での NIE では、各新聞社が発行している紙媒体の新聞や Web ニュースを使用している。しかし、これらの記事には、児童が学習していない漢字や理解することが困難な表現などが使われており、児童にとって難解である。そのため、日々発行される膨大な記事の中から児童自身が自分の興味や課題に適した記事を探し出すことが難しいという課題がある。また、教師にとっては、通常業務に加え、NIE に適した記事の選出や記事に関する教材研究などが必要になるため、業務負担が増加するという課題がある。したがって、記事の推薦や重要語の提示、難解な言葉の言い換えなど、児童に対する支援と NIE に活用しやすい記事や関連・補足資料の検索・選択など教師に対する支援が必要になる。そのため、新聞記事に関する読解支援[2]や児童に対する記事推薦[3]などの研究が行われている。

本研究では、NIE を実践する教師向けの支援に注目し、NIE に活用しやすい Web ニュース記事を地域学習 NIE の教材として推薦するシステムの構築を目的とする。NIE 教材として良質な記事を推薦することで、教師が記事を探す負担を軽減し、NIE を継続的に実践するための支援が可能と考える。本稿では、NIE に適した記事を判定する手法と Web ニュース推薦システムを提案する。

2. 地域学習 NIE に適した記事

本研究では、NIE に適した記事として、NIE ワークシートに着目する。NIE ワークシートは、新聞社が独自に作成し、Web 上に補助教材として公開しているものである。NIE ワークシートを公開している新聞社は 20 社 (内地域紙 17 社) しかなく[4]、全国の小学校をカバーできないが、ワークシートの作成には NIE の経験が豊富な教師の意見を取り入れていることから、教材として質の高い記事が選ばれていると考える。また、ワークシート内の記事を分析した結果、地元に関する生物・産業・文化に関する記事が多く選出されることを確認している[5]。

本稿では、NIE ワークシートに掲載された記事を学習データに利用し、SVM (Support Vector Machine) を用いて地域学習 NIE に適した記事を判定する手法を提案する。

Web News Recommendation System for Regional Study in NIE at Elementary Schools

[†] Seki Shinya, Graduate School of Engineering, Kagawa University

[‡] Kazuaki Ando, Faculty of Engineering and Design, Kagawa University

3. 地域学習 NIE に適したニュース記事の判定法

3.1. 方針

本研究では、NIE に適した記事の判定を文書分類による二値分類タスクとして捉え、SVM を用いて記事判定する手法を提案する。

SVM は、パターン識別モデルの 1 つで、ラベル付き訓練データからラベル間の境界 (超平面) を学習し、識別したいデータと超平面の位置関係によって判定する。入力データの形式はベクトルであり、本研究のように記事を判定する場合、事前に記事テキストから特徴 (素性) を抽出し、その素性を要素とする文書ベクトルを各記事に対して生成する必要がある。

本稿では、NIE に適した記事の判定に、記事内容とその読みやすさの 2 点が重要であると仮定し、2 種類の文書ベクトルと、そのベクトルに組み合わせる 5 種類の素性 (1. 形態素の意味カテゴリ, 2. 地域特徴語, 3. リーダビリティスコア, 4. 学習済み漢字割合, 5. 語彙レベル割合) について検討する。以降、ベースになる文書ベクトルと追加する各素性について説明する。

3.2. 文書ベクトル (BoW, W2V)

文書分類タスクでは、一般的に文書内の単語を基に文書ベクトルを生成する。本稿では、BoW (Bag of Words) と W2V (Word2Vec) による、2 種類の文書ベクトルを用いる。W2V には学習済みモデル[6]を用いる。

3.3. 形態素の意味カテゴリ (C)

W2V は、形態素の意味を考慮したベクトルであるが、BoW は形態素の出現のみに着目しているため、意味を考慮したベクトルではない。そこで、形態素の意味カテゴリに着目する。意味カテゴリの取得には、形態素解析器 Juman++ の解析結果に含まれるカテゴリ情報 (22 種) と、日本語語彙大系[7]に含まれる体言カテゴリ情報 (2,840 種) を用い、それらを素性とする。なお、素性値には、意味カテゴリごとの記事中での出現割合を用いる。

3.4. 地域特徴語 (W)

地域特有の物事に関する記事を分類するため、地域特徴語を用いる。本稿では、ワークシートの記事分析結果に基づき、都道府県ごとの「特産品」「祭事」「文化財」「工芸品」に関する特徴語を Web 上から収集し、地域特徴語辞書を生成し、素性に利用する。そして、上記の 4 カテゴリ毎の出現語数を素性値に用いる。

3.5. リーダビリティスコア (R)

記事の読みやすさを測る指標として、リーダビリティに着目する。本稿では、李らが提案したリーダビリティ公式[8]によって算出されるリーダビリティスコアを素性値に利用する。なお、リーダビリティスコアは、難解なテキストほど低くなる。

3.6. 学習済み漢字割合 (K)

李らのリーダビリティスコアは、品詞や語種といった形態素情報のみから読みやすさを測る指標であり、小学生が対象という観点を考慮したものではない。そこで、小学生に対する文字単位での難易度として、小学生が学

