2N-07

# Webニュースを補足する画像コンテンツの推薦に向けた画像周辺テキストの解説性に基づくスコアリング手法

小栗 太樹† 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

近年，小学校において，新聞を教材として活用する教育NIE（Newspaper in Education）が実施されている．NIEの実践報告書[1]によると，新聞記事を選ぶこと，読むことなどを通じて，読解力の向上や自己判断力などを養うことができると報告されている．しかし，一般の新聞記事やWebニュース記事は，子供向けに書かれておらず，また，内容を理解・補足するための図や写真などもほとんど付与されていないため，NIE実践時には，小学生が読めない，理解できないなどの問題がある．また，NIEを実践する教師においても，新聞記事を探す労力に加え，記事内容を補足する資料の準備にもさらに時間を要する問題が生じている．

そこで本研究では，教師が選択したWebニュース記事に対して，記事内容を補足する画像コンテンツをWeb上から検索して，提示するシステムの構築を目的とする[2,3]．提案するシステムは，Webニュース記事から抽出した重要語あるいは教師が入力した重要語に基づいてクエリを生成し，Web上から画像を検索する．検索した画像を，表やグラフなどの画像タイプ別に分類[3]し，ニュース記事の補足資料として適した並びにリランキングして提示する．

先行研究[4]では，検索画像のリランキングの初期検討として，画像の周辺テキストを用いてスコアリングを行った．その結果，周辺テキストでの用語説明文の出現回数から解説性を評価する指標がスコアリングに有効であることを確認した．本稿では，解説性に着目したスコアリングの実現に向けて，画像周辺テキストの解説性を判定する手法について検討する．

## 2. 関連研究

本研究の先行研究として，村田ら[2]は，画像検索クエリ生成法と画像周辺テキストを用いた検索画像のスコアリング法を提案した．村田らの手法では，ニュース記事とWikipedia記事，小学校の教科書に含まれる図表キャプションからクエリを自動生成する．そして，生成したクエリで検索された各画像の周辺テキストとニュース記事に対して，TF-IDFを重みとするベクトルを生成し，cosine類似度により，各画像をスコアリングする．

近藤ら[5]は，与えられたテキストから重要語を抽出し，外部APIを利用することで，対象テキストに関連する動画やブログ等のコンテンツを推薦する手法を提案した．近藤らの手法は，一般ユーザに対して幅広い内容のコンテンツを網羅的に検索・推薦することを目指したものである．小学校のNIEで利用することを考えた場合，多数の幅広いコンテンツを提示するより，記事の内容に関連する質の高いコンテンツを提示する必要がある．

Scoring Method Based on Expository Text around Image for Recommendation of Image Contents Supplementing Web News

† Taiki Oguri, Graduate School of Engineering, Kagawa University
‡ Kazuaki Ando, Factory of Engineering and Design, Kagawa University

## 3. 補足画像提示システムの概要

以下に，提案システムの処理手順を示す．

STEP 1：教師が選択したWebニュース記事に対して，重要語を抽出，または，教師が重要語を直接入力
STEP 2：重要語を基に，ニュース記事とWikipedia記事，教科書のキャプションからクエリを生成
STEP 3：画像検索APIを用いて画像を検索
STEP 4：ニュース記事と画像周辺テキスト，画像特徴などを基に，画像タイプの分類とスコアリング
STEP 5：検索画像を，総合ランキングを提示，またはタイプ別にリランキングして提示

以降，本稿では，STEP 4の画像スコアリングの実現に向けて，画像周辺テキストの解説性を判定する手法について検討する．

## 4. 画像周辺テキストの解説性判定手法の検討

NIE実践教師は，ニュース記事の補足資料として，仕組みを説明する図や，データをまとめたグラフ・表などを求める傾向がある．このような画像の周辺には，用語や事象・現象などを解説するテキストが存在する可能性が高いと考えられる．また，先行研究[4]では，周辺テキストから画像をスコアリングする妥当性および周辺テキストの解説性がスコアリングに有効であることを確認している．

本稿では，以下に定義する解説文タイプのいずれかの特徴を含む画像周辺テキストを解説性ありと判定する．

① 用語説明文
用語の意味や事象・現象などの仕組みを解説する文
② 比較文
複数の事柄を比較し，その違いなどを解説する文
③ データ解説文
統計データに対して解説する文

本稿では，画像周辺テキストの解説性の有無を，SVM（Support Vector Machine）で判定する手法を検討する．

### 4.1 データセット

Web上の画像および画像周辺テキストを収集してデータセットを構築する．無作為に選択したニュース記事の重要語を人手で抽出し，重要語を検索クエリとして，画像検索APIを用いて画像およびWebページを収集する．そこから，HTML構造を用いて周辺テキストを抽出する．事前調査の結果を基に，HTML内の対象画像のimgタグから前後にそれぞれ1タグ内に記載されている日本語テキストを抽出する．これを文字数が500を超えるまで続ける．またtitleタグ，対象画像の属性も周辺テキストとして利用する．

収集した画像周辺テキストに対して，解説性の有無を人手で判定し，正解ラベルを付与する．最終的に，400枚の

画像および周辺テキストからなるデータセットを構築した．データセット中，解説文と判定したものは205件である．

### 4.2 ベースライン素性

SVMのベースラインの素性として，文章分類タスクにおいて一般的なbag-of-wordsとword2vecを素性として用いる．

● bag-of-words (BoW) 素性

画像周辺テキスト中の名詞，動詞，形容詞，副詞のbag-of-wordsを素性として用いる．データセットにおいて，出現テキストが3%以下 90%以上の単語はノイズとして除去する．また，bag-of-wordsに対して，TFおよびTF-IDFで重み付けしたベクトルも素性として利用する．

● word2vec (W2V) 素性

単語の分散表現であるword2vec[6]を素性として用いる．word2vecにおいては，意味の近い単語から生成されたベクトルは類似したベクトルになることが期待される．本稿では，画像周辺テキストに出現する名詞，動詞，形容詞，副詞のword2vecベクトルを加算し，単語数で割った値を素性値とする．word2vecの学習には，2019年5月時点の日本語Wikipedia記事全文を用い，次元数は200とする．

### 4.3 解説性に基づく素性

桜井らの手法[7]を応用し，解説文で使われる特徴的な表現に着目した素性を提案する．桜井らは，Web上から用語説明文を収集するために，用語説明文に使われる表現を43パターンで整理した．この43パターンは，本研究における解説文の定義①で使われる表現が中心である．よって，定義②③の解説文で使われやすいと考えられる74表現を，データセットや実例などを参考に抽出した．表1にその一部を示す．さらに，これらに出現する82単語の類義語を日本語WordNet[8]から298語取得して語彙拡張した．最終的に，解説性に基づく415素性を利用する．なお，素性値は，各パターン・表現がテキスト内で出現する頻度とする．

表1. 追加した表現の例

| | |
|---|---|
| ② 比較文 | 違い，比べて，異なる，<br>分類，比較，対し，種類 |
| ③ データ解説文 | 平均，推移，傾向，調査<br>%，増加，減少，上昇， |

## 5. 評価実験

### 5.1 実験設定

提案した判定手法の性能を確認するため，評価実験を行う．形態素解析にはMeCabを用いる．SVMのハイパーパラメータは，グリッドサーチによりチューニングする．適合率，再現率，F値を評価尺度として，作成したデータセットを用いて，10分割交差検証で各素性に基づく結果を評価する．まず，ベースライン素性のみを用いた性能を確認し，その後，提案素性およびベースライン素性と提案素性を組み合わせた場合の性能を比較する．

### 5.2 評価結果

#### 5.2.1 ベースライン素性の評価結果

ベースライン素性を用いた実験結果を表2に示す．表2から，W2Vを素性とした場合，どの指標においても最も高い値が確認できる．逆に，BoWを素性とした場合は，W2Vと比べて低い適合率となった．

表2. ベースライン素性を用いた実験結果

| | 適合率 | 再現率 | F値 |
|---|---|---|---|
| BoW | 0.66 | **0.75** | 0.70 |
| BoW×TF | 0.70 | 0.70 | 0.70 |
| BoW×TF-IDF | 0.66 | 0.74 | 0.70 |
| W2V | **0.72** | **0.75** | **0.73** |

#### 5.2.2 提案素性の評価結果

次に提案素性を用いた実験結果を表3に示す．表3に示すように提案素性のみでは，ベースライン素性と比較して高い性能は得られなかった．しかし，ベースライン素性と組み合わせることで性能が向上した．特に，word2vec素性と提案素性を用いることで，どの指標においても最高値を得ていることが確認できる．提案素性を用いることで，解説性文での特徴表現が正確に判定できたと考えられる．

表3. 提案素性を用いた実験結果

| | 適合率 | 再現率 | F値 |
|---|---|---|---|
| 提案素性 | 0.72 | 0.65 | 0.68 |
| 提案素性+BoW | 0.68 | 0.76 | 0.72 |
| 提案素性+BoW×TF | 0.75 | 0.72 | 0.72 |
| 提案素性+BoW×TF-IDF | 0.66 | 0.74 | 0.70 |
| 提案素性+W2V | 0.75 | 0.79 | 0.77 |

## 6. おわりに

本研究では，小学校教師がNIEの授業準備の負担を軽減することを目的に，Webニュース記事の内容を補足する画像コンテンツを提示するシステムの構築を進めている．本稿では，そのシステムで利用する検索画像のスコアリングの実現に向けて，画像周辺テキストの解説性をSVMで判定する手法を検討した．評価実験の結果，解説文で使われる特徴を素性に用いることで判定性能の向上を確認した．

今後は，エラー分析を行い，新たな素性の導入などによる判定性能の向上を目指す．最終的に，システムとして実装し，総合評価する．

### 謝辞

本研究の一部はJSPS科研費 19K12271の助成を受けて実施した．

### 参考文献

[1] NIE実践報告書，https://nie.jp/report/pamflet/

[2] 村田他，“小学校におけるNIEのためのWebニュース記事を補足する画像コンテンツの検索”，IPSJ2018，pp.437-438，2018.

[3] 小栗他，“小学校におけるNIEのためのWebニュースを補足する画像の分類”，IPSJ2019，pp.561-562，2019.

[4] 小栗他，“小学校におけるNIEのためのWebニュースを補足する画像のスコアリング手法の検討”，FIT2019，pp.319-320，2019.

[5] 近藤他，“重要語抽出を用いた外部APIからの関連コンテンツ推薦”，JSAI2010，1D2-1，2010.

[6] Mikolov, T., et al., “Efficient Estimation of Word Representations in Vector Space”, Proc. of ICLR2013, 2013.

[7] 桜井他，“ワールドワイドウェブを利用した用語説明文の自動生成”，情処学論，Vol43，No.5，pp.1470-1481，2002.

[8] 日本語WordNet，http://compling.hss.ntu.edu.sg/wnja/