5L-05

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

村田 真澄† 安藤 一秋‡

†香川大学大学院工学研究科 ‡香川大学工学部

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）が注目を集めている．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[1]．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，内容を理解するための図や写真などもほとんど付与されていない．そのため，NIE 実践時には，小学生が読めない，理解できないといった問題がある．子ども新聞のように，子ども向けに書かれた記事も存在するが，一般の新聞に比べて記事数が圧倒的に少なく，小学校の高学年向けの NIE では，ほとんど利用されていない．そのため，NIE を実践する教師は，時間をかけて新聞記事を選択した後，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といった問題が生じている．

本研究では，後者の問題に注目し，教師が選択した Web ニュース記事に対して，記事の局所的な内容を補足する画像コンテンツを検索して提示するシステムの構築を目的とする．

本稿では，まず，関連研究と既存サービスについて述べ，次に，画像コンテンツの検索手法について説明する．そして，画像コンテンツを検索するための 3 種類のクエリ生成法について述べた後，画像周辺テキストを利用した画像コンテンツのスコアリング手法について検討する．

2. 関連研究

近藤らは，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[2]を提案している．この手法では，テキストから抽出した数語の重要語を OR 連結してクエリを構成するため，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，多くのコンテンツを提示するより，質の高いコンテンツを提示する方が効率的である．

本研究では，既存研究を参考にしつつ，教師が選択した Web ニュース記事に対して，記事の局所的な内容を補足する画像コンテンツを提示する手法を実現する．

3. 画像コンテンツの検索手法

ニュース記事の内容を補足する有用な画像コンテンツは，Web 上に多数散在していると考えられる．そこで，本研究では，ニュース記事に対し，記事の局所的な内容を補足する画像コンテンツを検索する手法を提案する．

以下に，検索手法の処理手順を示す．

STEP 1：教師が選択したニュース記事に対して記事の内容を分析し，重要語を抽出する．

STEP 2：3 種類の検索クエリ生成法により，クエリを生成する．

STEP 3：生成した 3 種類のクエリと重要語のみのクエリを用いて画像検索 API で検索する．

STEP 4：ニュース記事と画像周辺テキストの類似度・難易度やクエリ情報などを基に，4 種類のクエリに対する検索画像をスコアリングする．

STEP 5：取得したすべての画像に対し，スコアでリランキングして提示する．

以降，本稿では，記事から得られた重要語に対し，STEP 2 の検索クエリを生成する方法を述べ，その後，STEP 4 の初期段階として，画像周辺テキストを利用したスコアリング手法について検討する．

4. 画像検索クエリの自動生成

記事全体の内容を補足する画像を検索する方法として，記事内のすべての重要語を AND で連結したクエリの利用が考えられる．しかし，ニュースは，日々，多様な内容で発信されるため，記事全体の内容を 1 枚で表現した画像が存在する可能性はほとんど期待できない．

そこで本研究では，記事全体の内容ではなく，局所的な内容の補足に焦点をあて，重要語に対する記事内での役割や属性，動作，変化などの情報を付与したクエリを生成することで，記事の局所的な内容を補足する画像コンテンツを検索することを目的とする．

画像コンテンツの検索には，ニュース記事本文，Wikipedia 記事本文，教科書キャプションを利用して生成したクエリを利用する．

まず，ニュース記事と Wikipedia 記事に対し，パターンマッチングによりクエリ生成する手法について述べる．重要語と“の”で連接して共起する語は，重要語に対する役割や属性などの関係を示す可能性が高いと考えられる．そこで，重要語と“の”で連接する共起語を用いて検索クエリを生成する．具体的には，“重要語＋の＋共起語”を検索クエリとする．また，記事本文において重要語を含む文節が係る文節の内，サ変系単語（サ変名詞とサ変動詞）から始まる文節は，重要語に対して何らかの動作・変化を与える関係にあると考えられる．この関係を基にクエリ生成することで，重要語に対して動作・変化のある画像が検索できる可能性がある．そこで，重要語とサ変系単語の係り受け関係から“重要語＋の＋サ変系単語”というクエリを生成する．

次に，教科書キャプションを利用してクエリ生成する手法について述べる．本手法では，小学校の教科書に含まれる図表に付与されているキャプションに注目する．“東京書籍 新編 新しい社会 6 上（2010 年）”内のキャ

Search for Image Contents to Supplement Web News
Masumi Murata†, Kazuaki Ando‡
†Kagawa University, Graduate School of Engineering
‡Kagawa University, Factory of Engineering

プションを調査し，キャプションの末尾に頻出する 18 種の末尾語を人手で選定した．その後，“重要語＋の＋末尾語”というクエリを生成し，その有効性を評価した結果，重要語と関係をもたない末尾語がノイズになることを確認した．そこで，重要語と末尾語の共起関係を調べて末尾語を選定する方式を採用する．ここで，共起語の判定には，Shimpson 係数を改良した係数を用いる．実験[3]により，Shimpson 係数の分子に“X の Y”でフレーズ検索をしたヒット数を用いることで，より共起性の強いクエリが選定できることを確認した．この結果より，共起度の高い上位 3 語の末尾語を重要語と組み合わせ，クエリを生成する．

5. 画像周辺テキストを利用したスコアリングの検討

5.1 スコアリング手法

画像検索 API の結果をニュース記事の内容に適した並びにリランキングするため，ニュース記事の本文と Web 上の画像の周辺テキストの類似度を基にスコアリングする手法について検討する．

画像周辺テキストの抽出には，HTML タグ構造を利用する．HTML 内の画像 URL が記述されている箇所より，前後にそれぞれ 2 つの p タグ，計 4 つの p タグに記載されているテキストを画像周辺テキストとして使用する．p タグが取得できない場合は，画像 URL が含まれている h タグ内のテキストを全て利用する．h タグも取得できない場合は，ページ内の全てのテキストを使用する．ニュース記事本文と画像周辺テキストとの類似度は，TF-IDF 値を基にした cosine 類似度で測定し，スコアリングする．

5.2 簡易評価

スコアリング手法の性能を確認するため，簡易評価を行う．評価には，読売新聞の 7 記事を使用し，各記事から 1 語ずつ人手で重要語を選択する．各重要語から 3 つのクエリを生成して画像検索した後，スコアリング結果を基にリランキングされた上位 10 件の画像を主観評価する．評価は，重要語に適した画像であるか，ニュース記事に適した画像であるか，わかりやすい画像であるかの 3 点を 5 段階（1 が適切で 5 が不適切）で評価する．なお，IDF には，読売新聞の Web ニュース 2016 年 5 月から現在までに至るおよそ 33,000 記事のデータを利用する．

評価結果として，図 1 にランキング順位と画像スコアの相関グラフを示す．ランキング上位の画像の内，いくつかの画像はノイズ画像が含まれていたが，全ての評価項目において負の相関がみられた．この結果から，画像へのスコアリング，リランキングが作用していると考えられる．

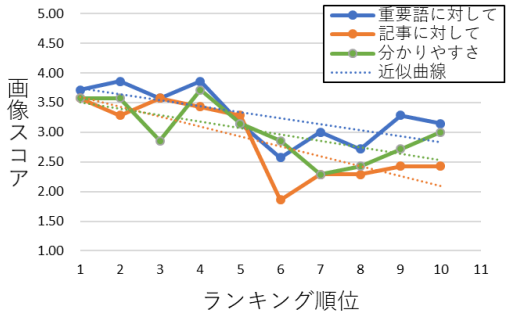

図 1. ランキングと画像スコアの相関

5.3 画像周辺テキストの取得範囲の影響

画像周辺テキストの取得文字数を変更しながら，リランキングに与える影響について調査する．具体的には，初期検討においては，画像タグから前後 2 個の p タグを周辺テキストとしていた．本調査では，閾値 x を超えるまで p タグを取得し，周辺テキストとするように変更した．閾値 x は 100～500 まで 100 刻みで変更する．また，画像タグを含む行のテキストと画像タグの alt 属性についても周辺テキストとして加えた．

5.1 の評価と同様，3 つの観点を 5 段階で評価する．なお，評価の際，ランキングの上位ほど重視すべきであると仮定し，順位に基づく重みづけ（1 位ならば 10 倍，2 位ならば 9 倍…10 位ならば 1 倍）を行う．

評価結果を表 1 に示す．表中の AVG は平均を意味する．表 1 から，文字数を 300 文字に設定した場合の AVG が最も高いことから，閾値は 300 文字が妥当であるといえる．5.1 の p タグを 4 個利用する手法は，AVG が最も低い．これは，取得テキストが 0 文字になる場合があるため，スコアが低くなってしまったと考えられる．

表 1. 周辺テキスト量によるリランキング

文字数	重要語に対して	記事に対して	分かりやすさ	AVG
100	3.39	2.63	2.91	2.98
200	3.38	2.58	2.95	2.97
300	3.47	2.66	3.09	3.07
400	3.43	2.56	3.02	3.00
500	3.35	2.68	2.86	2.96
pタグ4個	2.78	2.44	2.50	2.57

6. おわりに

本稿では，小学校での NIE に利用するため，ニュース記事の局所的な内容を補足する画像コンテンツを検索する手法を提案した．具体的には，画像コンテンツを検索するための 3 種類のクエリ生成法を提案した後，画像のスコアリング手法について検討した．評価の結果，画像周辺テキストの文字数として 300 文字が適切な閾値であることを確認した．

今後は，スコアリングに画像周辺テキストの難易度やクエリ生成法に対するパラメータを導入することで性能の向上を目指す．そして，最終的には，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 16K00478 の助成を受けて実施した．

参考文献

[1] NIE 教育に新聞を，http://nie.jp/
[2] 近藤光正，中辻真，田中明通，“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 論文集，1D2-1，pp.1-4，2010.
[3] 村田真澄，安藤一秋，“Web ニュース記事を補足する画像コンテンツの検索”，IPSJ2017 講演論文集，2K-06, pp.451-452，2017.
[4] 村田真澄，安藤一秋，“小学生のためのニュース記事を補足する画像コンテンツ検索用クエリの検討”，FIT2016 講演論文集，N-020, pp333-334，2016.