N-020

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

A Study on Query for Searching Image Contents to Supplement the Contents of Web News for Elementary School Students

村田 真澄† 安藤 一秋‡
Masumi Murata Kazuaki Ando

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ため，小学生が読めない，理解できないといった問題がある．子供新聞のように，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いため，小学校の高学年向け NIE ではほとんど利用されていない．

新聞記事の読解支援として，新聞記事内の難しい単語や表現を平易に言い換える方法や記事内容の理解を助けるための関連する情報を提示する方法などが考えられる．
そこで本研究では，後者に注目し，記事に関連する画像コンテンツを検索して，小学生に提示するニュース記事読解支援システムを提案する．本稿では，ニュース記事読解支援システムについて概説すると共に，記事の局所的な内容を補足する画像を Web 上から検索するためのクエリ生成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 で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，小学校での NIE 利用を想定し，網羅性より適合性の向上に焦点をあて，記事の局所的な内容に関連する画像を検索・提示するシステムの実現を目指す．

3. ニュース事読解支援システムの概要

ニュース記事の読解を支援するための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，ニュース記事推薦手法[3]により得られた記事群に対し，記事の局所的な内容に関連する画像コンテンツを提示する読解支援システムの実現を目指す．

†香川大学 大学院工学研究科 Kagawa University Graduate School of Engineering
‡香川大学 工学部 Kagawa University Factory of Engineering

以下に，システムの処理手順を示す．

1. 記事推薦手法で得られた記事群に対し，各記事の内容を分析し，重要語を抽出する．
2. 重要語等を基に検索クエリを生成する．
3. Web 画像検索 API を用いて関連画像を検索する．
4. ノイズ画像をフィルタリングし，画像周辺テキストなどの情報を基にリランキングした結果を提示する．

以下，本稿では，各記事から得られた重要語に対し，検索クエリを生成する方法について検討する．

4. 画像検索クエリの検討

記事全体の内容に関連する画像を検索するには，記事内のすべての重要語を AND で連結してクエリを構成すればよい．しかし，ニュース記事は，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あまり期待できない．そこで本稿では，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などの情報を付与したクエリを生成することで，重要語に関連する画像コンテンツの検索を最初の目標とする．

本稿では，記事内で重要語と“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仮定し，記事内の重要語と“の”で連接する共起語を用いて検索クエリを生成する手法について検討する．具体的には“記事の重要語＋の＋共起語”を検索クエリとして画像検索する．

クエリを構成するための共起語は，(1)ニュース記事内で重要語と“の”で共起する語；(2)重要語を見出しとする Wikipedia 記事内で重要語と“の”で共起する語；の 2 種類について検討する．さらに，教科書内の図表等に付与されたキャプションに注目し，(3)キャプションの末尾に頻出する語（仕組みや使い方，様子など）を共起語の代わりに用いる手法も検討する．

調査方法について説明する．検索エンジンの API の関係から画像検索には，Microsoft Bing を利用する．検索結果の上位 10 件について，検索クエリと関連のある画像でかつ，小学生でも理解できそうな画像か否かを人手で判定し，適合率で評価する．調査対象のニュース記事は，任意の 10 記事とする．各記事の重要語は，見出しとリードに共に含まれる語の内，任意の 1 語を利用する．

4.1 記事本文中の共起語の利用

記事本文から“重要語＋の＋名詞”なるパターンで共起語を抽出してクエリを生成する．調査結果を表 1 に示す．

10 件の新聞記事のうち，3 件の記事において共起語を抽出できた．これらの共起語からクエリを生成し，画像検索した結果，適合率は表 1 のようになった．共起語を抽出で

きた件数は 10 件中 3 件と少ないが，記事内の共起語を使うため，記事と直接関連のある画像を検索できた．

表 1　記事本文の共起語に対する評価結果

重要語	クエリ個数	適合率の最大値(%)	適合率の平均値(%)
温暖化対策	1	30.0	30.0
天然資源	2	60.0	40.0
オゾンホール	0		
天然記念物	0		
風力発電	0		
液状化	1	100.0	100.0
平均気温	0		
スーパーコンピューター	0		
ビッグデータ	0		
宇宙ステーション	0		

4.2　Wikipedia 内の共起語の利用

重要語を見出しとする Wikipedia 記事の本文において，重要語と“の”で共起する語を抽出してクエリを生成する．Wikipedia を利用することで，ニュース記事の本文から得ることができなかった共起語を取得できる可能性がある．調査結果を表 2 に示す．

表 2　Wikipedia 記事内の共起語に対する評価結果

重要語	クエリ個数	適合率の最大値(%)	適合率の平均値(%)
温暖化対策	0		
天然資源	2	50.0	25.0
オゾンホール	3	40.0	23.3
天然記念物	10	100.0	10.0
風力発電	15	100.0	50.7
液状化	4	70.0	30.0
平均気温	4	100.0	60.0
スーパーコンピューター	21	10.0	1.4
ビッグデータ	7	20.0	7.1
宇宙ステーション	0		

10 件中 8 件の記事について Wikipedia から共起語を抽出できた．Wikipedia から抽出できた共起語の数は重要語によって大きく異なった．また，重要語と完全一致する記事が Wikipedia に存在しない場合もあった．そこで，記事が存在しない場合は，人手で判断した Wikipedia の記事を使用した．結果として，Wikipedia 記事内から複数の共起語を抽出できた．生成したクエリで検索した結果，一部の重要語を除いて，多くの有意な画像を検索できた．しかし，記事と直接関係のない画像も一定数含まれていた．“スーパーコンピューター”や“ビッグデータ”などの重要語については，多くの共起語を抽出できたが，あまり有意な画像は検索できなかった．画像で表現することに向いていない共起語や小学生には難しい共起語が影響している．

4.3　教科書キャプションの末尾語の利用

まず，教科書“東京書籍 新編 新しい社会 6 上（2010 年）”内の図表や写真のキャプションの末尾語を調査した結果，キャプションの末尾には“使い方，生活，様子，産物，年表，関係，つくり，歩み，変化，内容，流れ，例，仕組み”等が使われていた．本調査では，末尾語として頻出する 13 語を利用する．調査結果を表 3 に示す．

表 3　科書キャプションの末尾語に対する評価結果

重要語	クエリ個数	適合率の最大値(%)	適合率の平均値(%)
温暖化対策	13	70.0	20.77
天然資源	13	40.0	10.77
オゾンホール	13	80.0	26.15
天然記念物	13	60.0	9.23
風力発電	13	100.0	33.08
液状化	13	90.0	21.54
平均気温	13	100.0	30.77
スーパーコンピューター	13	60.0	10.77
ビッグデータ	13	40.0	11.54
宇宙ステーション	13	60.0	16.15

すべての重要語に対して，何らかの画像を検索できた．しかし，重要語と末尾語を単純に“の”で連結してクエリ生成したため，通常では共起しない組み合わせになるクエリが存在した．このようなクエリに対しては，重要語と末尾語間に関係のない画像が多く検索された．適合率の最大値が示すように，一般的に共起すると考えられる組み合わせのクエリからは，有用な画像が検索できた．

4.4　考察

ニュース記事本文の共起語を用いた手法は，記事と直接関連のある画像を検索しやすいが，共起語を取得できない可能性がある．Wikipedia 記事内の共起語を用いた手法は，多くのクエリを生成できると共に妥当な画像も多数検索できた．教科書キャプションの末尾語を用いた手法は，ニュース記事とあまり関連のない画像も検索されるが，固定された語を利用することで，重要語に依存せずに関連画像を検索できた．記事との関連性を考慮すると，本文中の共起語，Wikipedia 記事内の共起語，キャプションの末尾語の順でよい結果が得られた．検索できる画像数の観点からは，逆の順位となる．本システムでは，記事に関連する画像の方が望ましいため，前者を採用する．

5. おわりに

本稿では，ニュース記事に関連する画像を追加提示することで，小学生に対する読解を支援するシステムを提案した．そして，画像を検索するためのクエリ生成法を検討した．今後は，共起語の選別方法や係り受け関係によるクエリ生成法について検討する．そして，ノイズ画像のフィルタリングと画像周辺テキストなどを利用したリランキング手法について検討し，システムを実装する．

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 論文集，1D2-1，pp.1-4，2010.

[3] 田中他，“Facebook での注目度に基づいた小学生のための Web ニュースランキング”，IEICE-ET2015-97，pp.21-26，2016.