研究指導 中澤 真 教授

テキストの難易度に基づく子供向け Web ページ判定手法の提案

鈴木 美羽

1. はじめに

インターネットの普及により子供がインターネットを利用することが増えている．青少年のインターネット利用環境実態調査[1]によると，小学生の多くはインターネットを日常的に利用しており，利用率は7歳では68.9％，12歳では92.8％となっている．また，教育現場での利用も進んでおり，NTTレゾナントによる調査では教員の80.7%が小学校の調べ学習において児童にインターネットを使用させていることが明らかになっている[2]．しかし，子供がインターネットを使用する場合にはアダルトコンテンツや暴力などに関する不適切な情報を閲覧したり，トラブルに巻き込まれたりなどの危険性への対応が必要になる．そこで，対策として子供向けの検索エンジンやフィルタリングサービスの提供，18歳未満の子供が使用するスマートフォンへのフィルタリングの導入の義務化などが実施されている[3]．

このように子供が安全にインターネットを利用できるようにする取り組みが行われているが，これらのサービスは不適切なコンテンツの閲覧を防止するだけであり，子供の理解力に適した難易度かどうかは考慮されていない．そのため，子供が学習などのためにWeb検索を行った際に難しいWebページばかりが表示されることも考えられ，結果として内容を理解できなかったり，取り組みそのものを放棄してしまったりなど効果的な学習を行えない可能性が生じる．

この問題に対して，Webページの難易度を推定して適切なものだけを子供に提示するための研究がいくつか行われている[4][5][6]．これらは難易度推定の特徴量として，Webページの画像の多さや，語りかけ口調の文末表現の量などを用いている．しかし，極端に低年齢層の子供向けページばかりが子供向けと判別されるなど，その精度には課題も多い．Webページが子供にとってわかりやすいか否かは，Webページのデザインや文体が子供向けらしいという表面的な部分よりも，文章の構造やその複雑さに左右されると考えられ，これが高い精度にならない原因と考えられる．

文章の複雑さに関連する特徴量を用いた研究としては，各学年で習う漢字の割合や各級の語彙などを特徴量として難易度を推定しているものがある[7][8]．しかし，特定の学年を対象とし，その学年までに習う文章表現のみで書かれている文章は教材以外では稀である．そのため，Webページなどの一般的な文章をこの手法で判別しても同様の推定精度は得られない恐れがある．

そこで本研究では，教科書や教材以外の一般的なWeb文章においても難易度推定ができる方法を確立するため，漢字の配当学年のような教科書特有の特徴量ではない新しい特徴量を模索する．さらに，これらの特徴量を用いて対象となる文章を子供向けか大人向けか判別した際に，これまでよりも高い判別精度を達成できることを明らかにする．

2. Web ページの難易度推定の現状

現在，情報活用能力の育成が推進されており，教育のICT化に向けた取り組みとして，教育現場への環境整備が進められている[9]．これにより，小学校ではインターネットを用いた調べ学習やネットリテラシーについての指導が行われており，子供がインターネットを利用することが一般化してきている．これに伴い，子供が安全にインターネットを利用するためのサービスやツールも必要になり，子供向けの検索エンジンではYahoo!きっず[1]，フィルタリングソフトではノートンファミリー[2]など様々なものが有償無償の別を問わず提供されている．しかし，既存のサービスではアダルトコンテンツや暴力などといった不適切なWebページをブロックしてくれるが，コンテンツ内で使用されている文章の複雑さや使用される漢字の難易度などが，子供の理解度に応じた適切なものであるか否かは考慮されない．小学生は検索結果の上位5件しか閲覧しない傾向にあるという結果[10]からも，子供には難しすぎるページを除外して検索結果を提示できることが望ましい．

そこで，Webページの難易度を推定し，子供が読むのに適切なWebページを判別する手法がいくつか研究されている[4][5][6]．この推定において重要なのが，難易度推定に必要なWebページの特徴量としてどのような情報を用いるかということである．これは，特徴量の選び方が推定精度に大きな影響を及ぼすからである．そこで次節では難易度推定のための特徴量について議論する．

3. 難易度推定に用いる特徴量

3.1 先行研究の特徴量と課題

岩田ら[6]の研究では，子供は文字よりも画像を好むという傾向を考慮して，難易度推定の特徴量としてWebページ内の画像やアニメーションの量を用いた．しかし，この特徴量では画像やアニメーションを多用した低年齢の子供向けのページや娯楽に関するページばかりが子供向けであると判別され，子供にも容易に読める難

1 https://kids.yahoo.co.jp/

2 https://family.norton.com/web/

易度の文章のページを子供向けではないと判別してしまう恐れがある.

一方, 文章量ではなく文章そのものに注目した特徴量として,「～しよう」,「～だよ」のように読者に語り掛けるような文末表現の割合を用いた研究がある[5]. しかし, 大人向けのブログでも親しみやすさを感じさせるために文語文ではなく口語文が使われることが多く, このような文末表現が出現しやすい. このため, 誤判別を招きやすく判別精度も36%にとどまっている. 文章に出現する単語の傾向に着目し,「小学生」や「キッズ」など特定の単語の有無を特徴量とした研究もある[4]. しかし, これらの単語は保護者や教師向けの文章にも多く出現する傾向にあるため誤判別してしまうことも多く, 実際に正解率も46.2%と低い精度になっている. いずれの研究も判別精度が低かったことから, 文末表現や特定のキーワードの有無といった特徴量では, 文章の内容的な難しさまで反映させることができないと考えられる.

これに対し, 教科書や言語学習用教材のように, 文章の内容的難易度が明確になっている文章を用いることで, 難易度推定に効果的な特徴量を選び出している研究もある. その一つに, 配当学年別の漢字の出現回数や, 品詞の出現回数, 字種[3]の出現比率を特徴量として用いた研究[7]がある. この研究では教科書の文章を低学年, 高学年, 中学生の3つに判別することを目的としており, 先の特徴量を用いることで適合率, 再現率ともに70%を超える高い精度を実現している. また, 日本語能力試験の各級で習う語彙の個数や文章の係り受けの関係を数値化したものを特徴量とした研究[8]でも正解率は72.2%と高い結果となっている. しかし, 前者の研究は教科書を, 後者は日本語能力試験の教材を使用しており, 漢字の配当表や各級の語彙の割合を使用すれば, 教科書や教材に対する判別精度が高くなるのはある意味当然である. 教材以外の一般的なWebページの文章では, 読み手の対象学年を意識し, 該当する学習済み漢字・語句・文法のみを用いて書かれていることは稀であるため, このような特徴量を使用した推定手法では十分な精度が得られない可能性が高い. 教科書や教材以外の文章にも応用可能な手法とするためには, 学年や級で明確に定められた特徴量ではないものを使用するべきである.

3.2 本研究で使用する特徴量

本研究では, 係り受け距離, 漢字使用率, 読点間の長さ, 文長, 抽象名詞の割合の6つを特徴量の候補とする.

係り受け距離は劉ら[8]でも使用されており, 単語がどれだけ離れた単語に係っているかを表している. 本研究では, 構文解析器CaboCha[4]で付与される文節番号を用いて「文節番号－係り先の文節番号」でその文節の係り受け距離を算出し, 特徴量として使用している. 係り受け距離が短いほど文の構成は簡素になり, 読みやすくなるはずである. そのため, 係り受け距離は難易度を推定する特徴量として適切であると考えられる.

次の特徴量候補は漢字使用率である. 漢字使用率は北内ら[7]でも使用されており,「文章中の漢字の文字数/文章全体の文字数」で算出される. 上級学年になるほど習得済み漢字の量が増えるため, 学年が上がるほど文中で使用される漢字の割合は一般的に高くなることが予想される. 当然, 大人向けの一般文章はさらに漢字使用率は高くなるはずであり, 難易度を推定する特徴量として適切であると考えた.

次の特徴量候補は読点間の距離である. この特徴量を難易度推定に用いた研究はまだない.「文章全体の文字数/文章中の句読点の数」で文章中の平均の読点間の距離が算出される. 文は読点が打たれたところで一区切りとなるため, 読点間の距離が小さいほど文が短いまとまりで構成されることになる. そのような文は簡潔で子供でも読みやすくなるため, 読点間の距離を難易度推定のための特徴量とすることは適切であると考えた.

次の特徴量候補は文長である. 文長は岩田ら[6]でも使用されており,「文章全体の文字数/文章中の句点の数」で文章中の平均の文長が算出される. 平均の文長が短いほど文章全体が簡潔であると予想できる. そのため, 文長は難易度を推定する特徴量として適切であると考えた.

最後の特徴量候補は抽象名詞の割合である. この特徴量を難易度推定に用いた事例はまだない. 抽象名詞の割合は,「文章中の名詞の数/文章中の抽象名詞の数」で算出される. ここでの抽象名詞とは現象や動作, 様子など実体のないものや概念を表す名詞である.「山」や「犬」のような一般的な名詞に比べ, 抽象名詞が多くなるほど文章はわかりにくいと考えられ, 抽象名詞の割合が大きくなるほど難易度は高くなると予想できる.

本研究ではこれら6つの特徴量の候補からどれを用いるべきかを検証し, 最終的に文章の難易度推定の精度向上を目指す.

4. テキストの分析と結果

4.1 分析方法

本研究では, 小学校の国語の教科書に掲載されている文章を各学年6つずつと, 日本経済新聞電子版のITと経済の2分野の記事を6つずつの計48件を抽出した. まず, 抽出したデータをトレーニングデータと評価用データそれぞれ24件ずつに分ける. そして, トレーニングデータを用いて特徴量の適切さの確認, 難易度の判別ルールの構築, 評価用データを用いた精度の検証をする.

子供向けの文章として使用した教科書は東京書籍の2019年度のものである. 1年生の教科書の上巻は口語詩が多かったり, 長さが文章として使用するには不十分であったりするため下巻を使用した. よって, 条件を揃えるために教科書が上下巻に分かれている学年は下巻を使用することとした.

なお, 本研究では大人向けの文章の学年を12として

[3] 漢字, ひらがな, カタカナ, アルファベット

[4] https://taku910.github.io/cabocha/

いる．これは，18歳が小学1年生から数えて12番目となるためである．また，後述する重回帰分析においても，大人向けの文章の学年設定値を7から13で算出した際に，12の場合の決定係数が最大となったことも理由の一つである．

4.2 係り受け距離

文節ごとの係り受け距離を求めた後に，文章全体に対する平均値を最終的な係り受け距離として特徴量として用いる．この平均値の算出の際，本研究では係り受け距離1のものを含めた場合と除いた場合の二つを特徴量の候補とした．文章が長くなるほど距離1となる隣の文節への係り受けが多くなるため，いずれの平均値も小さくなってしまい学年間の差が現れにくいという問題が生じたからである．

表 1と表 2は，学年ごとに3種類の文章(トレーニングデータ)に対して係り受け距離を算出した結果である．

表 1 係り受け距離(距離1を含めた場合)

	1年生	2年生	3年生	4年生	5年生	6年生	大人（IT）	大人（経済）
	1.96	2.06	1.97	2.21	2.57	2.28	1.92	2.18
	1.83	2.47	2.71	2.38	2.44	2.03	2.40	2.36
	1.92	2.07	2.00	1.98	2.16	2.28	2.49	1.96
平均	1.90	2.20	2.23	2.19	2.39	2.20	2.27	2.17

表 2 係り受け距離(距離1を除いた場合)

	1年生	2年生	3年生	4年生	5年生	6年生	大人（IT）	大人（経済）
	3.21	3.42	3.47	4.54	5.45	3.60	3.75	4.60
	3.57	4.54	4.20	5.30	4.38	3.53	5.05	4.62
	3.94	3.53	3.32	3.53	3.87	5.07	4.79	3.69
平均	3.57	3.83	3.66	4.45	4.57	4.07	4.53	4.30

表 1では学年による差が小さいが，表 2では高学年になるほど係り受けの距離が大きくなる傾向があることがわかる．学年と係り受け距離の相関をそれぞれ求めたところ，距離1を含めた場合の相関係数は約0.19，距離1を除いた場合は約0.36となり，後者のほうが学年との相関が強いことがわかった．それゆえ，係り受け距離(1を除いた場合)が特徴量として適切であると考えられる．

4.3 漢字使用率

漢字使用率の算出には総合文字数カウンター[5]を使用した．その結果を表 3に示す．この結果から，学年が大きくなるとともに漢字使用率が高くなる傾向にあることが読み取れる．これは，小学校の教科書はその学年までに習う漢字しか使用しないが，大人向けの文章ではそのような配慮はされていないことによるものと考えられる．学年と漢字使用率の相関を求めたところ，相関係数は0.93と高い数値であった．そのため学年と漢字使用率には強い相関があり，特徴量として適切であるといえる．

表 3 漢字使用率

	1年生	2年生	3年生	4年生	5年生	6年生	大人（IT）	大人（経済）
	1.18%	7.51%	11.16%	18.94%	27.20%	16.03%	34.90%	48.66%
	4.77%	11.15%	17.24%	21.01%	15.66%	21.88%	34.41%	41.87%
	7.41%	10.74%	9.04%	15.83%	19.47%	27.95%	32.46%	33.51%
平均	4.45%	9.80%	12.48%	18.59%	20.78%	21.95%	33.92%	41.35%

4.4 読点間の長さ

表 4**エラー！参照元が見つかりません。**は学年別に各文書の読点間の距離を分析した一覧表である．この結果から，小学生向けの文章では読点間の長さの学年による差がそれほど大きくないことがわかる．低学年ほど漢字の量が少ない(表 3)ことから，低学年の文章ではひらがなが多くなり，結果として1文あたりの文字数が増えてしまい学年間の差を小さくしていると考えられる．しかし，小学生向けの文章と大人向けの文章では読点間の長さの差は大きい．また，学年と読点間の長さの相関を求めたところ，相関係数は0.82となり，相関が認められた．そのため，読点間の距離は特徴量として適切であるといえる．

表 4 読点間の距離

	1年生	2年生	3年生	4年生	5年生	6年生	大人（IT）	大人（経済）
	9.20	8.05	9.89	12.58	14.12	11.04	24.41	21.76
	10.61	11.69	11.31	23.46	9.66	13.02	17.69	26.67
	11.46	9.96	10.21	10.86	11.09	12.42	22.06	21.41
平均	10.42	9.90	10.47	15.63	11.62	12.16	21.39	23.28

4.5 文長

表 5は学年別に各文書の文長を分析した一覧表である．この結果から，小学生向けの文章の多くが大人向けの文章より文長が短いことがわかる．子供向けの文章でも文長が40を超えている文章が2つあるが，これらは評論文である．評論文は物事の説明が多いため1文の長さが長くなったと考えられる．学年と文長の相関を求めたところ，相関係数は約0.68となり，相関が認められた．そのため，文長は特徴量として適切であるといえる．

表 5 文長

	1年生	2年生	3年生	4年生	5年生	6年生	大人（IT）	大人（経済）
	17.85	19.11	19.57	38.76	47.52	28.26	42.54	39.17
	23.50	34.64	33.94	46.92	32.40	26.18	59.50	48.00
	29.28	18.77	20.84	21.31	26.39	25.40	44.13	35.35
平均	23.54	24.17	24.78	35.66	35.44	26.61	48.72	40.84

4.6 抽象名詞の割合

文章中の名詞を抽象名詞とそれ以外の名詞に分類し，抽象名詞の割合を算出する．名詞の分類には日本語形態素解析システムJUMAN++[6]を使用する．JUMAN++では名詞に全22種の意味カテゴリが付与され，抽象名詞カテゴリとラベル付けされた名詞を抽象名詞とする．

表 6は抽象名詞の割合の一覧である．この結果から，学年が上がるほど抽象名詞の割合が大きくなることと，大人向けの文章は子供向けの文章と比べて抽象名詞の割合に大きな差があることが読み取れる．文章の内容を見ると，低学年向けの文章では誰が，何を，どうしたというような文章中で実際に起きていることが書かれていることが多いのに対して，高学年になるほど場の様子や登場人物の心情など抽象的なことも多く書かれるようになっていた．そのため，学年が上がるにつれて抽象名詞の割合が大きくなったと考えられる．学年と抽象名詞の割合の相関を求めたところ，相関係数は0.89となり，強い相関が認められた．そのため，抽象名詞の割合

[5] http://attosoft.info/tools/character-counter/

[6] http://nlp.ist.i.kyoto-u.ac.jp/index.php

は特徴量として適切であるといえる.

表 6 抽象名詞の割合

	1年生	2年生	3年生	4年生	5年生	6年生	大人（IT）	大人（経済）
	7.27%	13.79%	28.09%	12.00%	20.71%	39.85%	56.61%	67.12%
	16.67%	17.76%	31.43%	16.46%	16.94%	27.18%	64.72%	56.25%
	1.49%	10.67%	14.80%	13.24%	42.69%	21.02%	63.99%	57.82%
平均	8.48%	14.07%	24.77%	13.90%	26.78%	29.35%	61.77%	60.40%

5. 難易度判別式の算出と検証

5.1 重回帰分析による判別式の算出

本研究では，判別式を重回帰分析によって求める．ここで，目的変数は学年，説明変数は各特徴量とした．yは学年，cは定数項，x_iは各特徴量，a_iは回帰係数，6は特徴量の数である．以下が判別式である．

$$y = c + \sum_{i=1}^{6} a_i x_i$$

重回帰分析では特徴量同士の相関が強いとマルチコが発生する[11]ため，その場合は特徴量候補を削減することにした．まず，他の特徴量との相関が強い文長を候補から外した．この状態で重回帰分析をしたところ，係り受け距離（距離1を含めた場合）でマルチコが発生した．そのため係り受け距離（距離1を含めた場合）も候補から外して重回帰分析を行った．最終的にマルチコが発生しない状態とするために，係り受け距離（1を除いた場合），漢字使用率，読点間の長さ，抽象名詞の割合の4つを特徴量として用いる．これらを説明変数x_1, x_2, x_3, x_4としてトレーニングデータを標準化して重回帰分析した判別式を以下に示す．

$y = -1.80 + 0.20x_1 + 18.61x_2 + 0.07x_3 + 5.98x_4$

この式を用いて，$y \leqq 6$のとき子供向けの文章であると判別する．

5.2 判別精度の検証

判別式の精度を検証するためK分割交差検証を行う[12]．この検証方法では，データをトレーニングデータと評価用データに分割し，K回検証を行うものである．本研究では，K=10とし，子供向けの文章36個と大人向けの文章12個から各学年3つずつランダムにトレーニングデータと評価用データに分け，10通りの組み合わせで判別式の算出と検証を行う．

エラー! 参照元が見つかりません。は10通りの実験に対する判別の正解率を示したものである．

表 7 分割交差検証の結果

組み合わせ	正解率	組み合わせ	正解率
1	83.33%	6	83.33%
2	95.83%	7	87.50%
3	83.33%	8	83.33%
4	95.83%	9	87.50%
5	87.50%	10	83.33%

正解率が最も低いのは83.33%，正解率が最も高いのは95.83%であり，全ての組み合わせで正解率が80%を超える結果となった．また，平均正解率は87.08%となった．日本語の教材を用いて難易度推定を行った研究[8]の正解率は72.2%であり，本研究ではそれより高い精度での判別ができた．以上の結果より，本研究で提案する手法は有効であるといえる．

6. むすび

本研究では，小学校の国語の教科書に掲載されている文章と日本経済新聞電子版の記事を分析し，4つの特徴量から学年を予測し，難易度を判別する手法を提案した．文章の難易度推定に抽象名詞の割合を使用した研究は未だ存在していなかったが，検証の結果，正解率は全ての組み合わせで80%を超え，提案手法は有効であることが確認できた．この手法で用いた特徴量は文章の複雑さに関するもののみであることと，高い判別精度を持つことから子供向けのWebページの判別も可能であると考えられる．

今後の課題としては，誤判別しやすい高学年向けの評論文への対応である．これは，評論文は学年が上がるにつれて詳しい説明が増えるため文章が複雑になり，大人向けの文章との差が小さくなったことによると考えられる．よって，今後は文章データを増やして分析し，高学年向けの評論文と大人向けの文章の相違点を明らかにする必要がある．

参考文献・URL

[1] 内閣府，“第 2 部 調査の結果“，平成 31 年度 青少年のインターネット利用環境実態調査， pp. 132-351，2019，https://www8.cao.go.jp/youth/youth-harm/chousa/h30/net-jittai/pdf-index.html，（参照 2020-1-29）

[2] NTT コム リサーチ，“小学校教員を対象とした公務に関するアンケート”，https://research.nttcoms.com/database/data/001486/，（参照 2020-1-29）

[3] 内閣府，“第 4 章 第 3 節 子供・若者を取り巻く有害環境等への対応“平成 30 年度版 子供・若者白書（全体版），pp. 163-168，2018，https://www8.cao.go.jp/youth/whitepaper/h30honpen/pdf_index.html，（参照 2020-2-7）

[4] 泉川 洸一郎，安藤 千秋，“子供向け Web サイト収集のためのクローリング手法の検討”，FIT 講演論文集 Vol,14 No. 2，pp. 231-232，2015

[5] 佐藤 倫太郎ら， “子供 Web コーパス構築のための子供向けページ判定法”，情報処理学会第 80 回全国大会講演論文集，pp. 445-446，2018

[6] 岩田 麻佑ら，“子供による Web 検索のための検索結果リランク手法“，情報処理学会論文誌 Vol. 52 No. 3，pp. 1055-1068，2011

[7] 北内 啓ら，“文書特徴を利用した教育コンテンツの難易度判定”，情報処理学会第 64 回全国大会講演論文集，pp. 13-14，2002

[8] 劉 志宇，内田 理，“日本語を学習する外国人を対象とした日本語テキスト難易度判定手法”，研究報告自然言語処理 Vol. 2012-IFAT-105 No. 11，pp. 1-5，2012

[9] 文部科学省，“第 2 部 第 11 章 ICT の活用の推進”，平成 30 年度 文部科学白書，pp. 386-397，2019

[10] Bilal，D. “Children’s Use of the Yahooligane! Web Search Engine“，Journal of the American society for information science Vol. 51，No.7，pp.646-665，2000

[11] 管 民郎， Excel で学ぶ多変量解析入門，オーム社，2013

[12] 高木 章光，鈴木 英太，最新データサイエンスがよ

～くわかる本，秀和システム，2019