文書特徴を利用した教育コンテンツの難易度判定

2X−01

北内 啓† 高木 徹† 山本 健一郎‡
†株式会社NTTデータ ‡通信・放送機構

1 はじめに

近年、ネットワーク上に様々な教育コンテンツが増加し、教育現場で学習の教材として利用されている。しかし、コンテンツを利用する際、分類情報が付与されている場合が少なく、目的に合ったコンテンツを容易に見つけ出せないという問題がある。そのため、コンテンツに分類情報を付与することが課題となる。特に教育コンテンツの場合、内容だけではなく、難易度によっても分類する必要がある。例えば、植物の光合成を小学校と中学校の両方で学習するというように、同じ内容を異なる学年で学習することがある。したがって、コンテンツの難易度を判定することが重要である。

本研究では、文書中のテキスト情報に関する特徴を用いることによって、教育コンテンツの難易度を自動的に判定する手法を検証する。小学校低学年、小学校高学年、中学校の参考書を分析し、難易度判定に有効な文書特徴について検討する。文書特徴を用いて難易度判定を行う評価実験を行い、難易度判定の性能を評価する。

2 教育コンテンツの特徴分析

教育コンテンツの難易度判定に用いる文書特徴を分析する。文書特徴として、学年別漢字配当表の漢字の出現比率、品詞の出現比率、字種の出現比率を利用することにした。コンテンツとしては、以下の学習参考書のデータを使用した。

・小学校低学年: 1〜3年算数
・小学校高学年: 4〜6年算数、理科、社会
・中学: 1〜3年理科、歴史

これらのデータを「参考書データ」とよぶ。また、本研究で難易度を判定する低学年／高学年／中学の3段階を「レベル」とよぶ。

各文書特徴の分析結果についてそれぞれ説明する。

2.1 学年別漢字配当表の漢字の出現比率

文書の難易度は、文書中に出現する漢字によって変わると考えられる。そこで、参考書データ中に出現する漢字を、小学校学習指導要領[1)]の学年別漢字配当表を用いて分析する。学年別漢字配当表は、小学校で学習する学習漢字1006字を学年別に配当したものである。

各レベル・学科の参考書データに含まれる漢字について、配当表の低学年の漢字、高学年の漢字、配当表にないその他の漢字それぞれの延べ出現回数比率を集計した結果を**表1**に示す。

表1から、参考書のレベルが上がるにしたがって、低学年の漢字の比率が低下し高学年の漢字の比率が増加しており、配当表の漢字の比率に明確な違いがあることが分かる。

表1: 参考書データ中の漢字配当表の漢字の延べ出現回数比率

レベル・学科	低学年	高学年	その他
低学年算数	0.972	0.027	0.001
高学年算数	0.789	0.203	0.009
高学年理科	0.771	0.224	0.004
高学年社会	0.672	0.270	0.058
中学理科	0.624	0.299	0.077
中学歴史	0.552	0.334	0.113

2.2 品詞の出現比率

文書中の単語の品詞の出現比率を文書特徴とすることを検討する。参考書データに対し日本語形態素解析システム「茶筌」を用いて形態素解析を行い、一般名詞、サ変名詞、固有名詞、自立動詞の延べ出現回数の比率をレベルごとに集計した。なお、参考書データにはひらがなを含む単語など、茶筌の形態素辞書に登録されていない単語が数多く使われているため、それらの単語を人手で形態素辞書に登録し、形態素解析ができるようにした。

集計結果を**表2**に示す。表2から、一般名詞とサ変名詞については、高学年、中学とレベルが上がるにしたがって比率が増加する傾向にあることが分かる。また、固有名詞は高学年社会と中学歴史で比率が高い。ただし、自立動詞の比率はレベルとの相関が低い。

表2: 参考書データ中の単語の品詞の延べ出現回数の比率

レベル・学科	一般名詞	サ変名詞	固有名詞	自立動詞	その他
低学年算数	0.091	0.013	0.002	0.079	0.815
高学年算数	0.139	0.020	0.003	0.077	0.761
高学年理科	0.178	0.015	0.003	0.112	0.692
高学年社会	0.172	0.035	0.030	0.093	0.670
中学理科	0.195	0.044	0.004	0.093	0.664
中学歴史	0.164	0.061	0.060	0.079	0.636

2.3 字種の出現比率

文書中の字種の出現比率を文書特徴とすることを検討する。参考書データ中の文字について、漢字、ひらがな、カタカナの延べ出現回数比率をレベルごとに集計した結果を**表3**に示す。表3から、レベルが上がるにした

Difficulty Judging of Educational Contents using Document Features
Akira KITAUCHI†, Toru TAKAKI† and Ken'ichiro YAMAMOTO‡
†NTT DATA CORPORATION
‡Telecommunications Advancement Organization of Japan

がって漢字とカタカナの比率が上がるのに対し、ひらがなの比率は下がる傾向にあることが分かる。

表 3: 参考書データ中の字種の延べ出現回数の比率

レベル・学科	漢字	ひらがな	カタカナ	その他
低学年算数	0.120	0.533	0.014	0.333
高学年算数	0.237	0.480	0.021	0.262
高学年理科	0.208	0.614	0.048	0.130
高学年社会	0.339	0.488	0.031	0.142
中学理科	0.333	0.447	0.053	0.167
中学歴史	0.433	0.336	0.065	0.166

3 実験

2 章で分析した文書特徴を用いて、参考書データの難易度判定の評価実験を行う。

3.1 実験条件

実験データには前述の参考書データを用いた。参考書の節をひとつの文書とした。各学年・学科の参考書データ (651 文書) の一部をテストデータとし、cross validation によって精度を測定した。

文書ごとに漢字配当表の漢字の出現比率、品詞の出現比率、字種の出現比率の 3 種類の文書特徴を抽出し、各文書特徴の素性値を要素とする特徴ベクトルによって文書を表現した。小学校低学年、小学校高学年、中学校の参考書データをそれぞれ特徴ベクトルの集合と考え、テストデータを小学校低学年、小学校高学年、中学校の 3 カテゴリに分類することによって難易度を判定した。各文書特徴の素性値とその素性数は以下の通りである。

漢字配当表 漢字配当表の各学年の漢字の延べ出現回数比率を素性値とし、小学 1〜6 年とその他の 7 個を用いた。

品詞の出現比率 文書中の単語の品詞の延べ出現回数比率を素性値とし、一般名詞、サ変名詞、固有名詞の 3 個を用いた。表 2 において自立動詞の分布は難易度を反映していないと考え、素性からは除外した。

字種の出現比率 文書中の字種の延べ出現回数比率を素性値とし、漢字、ひらがなの 2 個を用いた。表 3 においてカタカナの比率は漢字やひらがなに比べてかなり低く、レベル間の比率の差は誤差の範囲であると考え、素性からは除外した。

3 種類の文書特徴それぞれを用いたものと、複数の文書特徴を組み合わせた「字種＋品詞」「配当表＋字種＋品詞」の 2 種類について難易度判定の精度を評価した。分類アルゴリズムには SVM (Support Vector Machine)[2]を用いた。評価尺度として、適合率と再現率を用いた。カテゴリ c に対する適合率 Pc, 再現率 Rc はそれぞれ以下の式によって算出される。

Pc = カテゴリ c に正しく分類された文書数 / カテゴリ c に分類された文書数

Rc = カテゴリ c に正しく分類された文書数 / カテゴリ c に属する文書数

3.2 実験結果と考察

各評価項目における、レベルごとの適合率と再現率を**表 4** に示す。文書特徴を単独で使用した場合、全体的には配当表の精度が高く、特に低学年での精度が高い。配当表の漢字は低学年と高学年で明確な違いがあるのに対し、品詞と字種はその差が小さいことが精度の差となって表れたと考えている。

字種と品詞を組み合わせた場合は、字種や品詞単独の場合よりも高い精度を示した。字種や品詞単独では、学科によって精度の高い学科と低い学科があるが、両者を組み合わせることにより、各学科において精度が高い方の特徴が重点的に利用されるため、全体として精度が向上したと考えている。

各漢字を人手で各学年に割り当てた配当表と、テキスト情報の比率のみを用いた「品詞＋字種」を比較すると、低学年では配当表の精度が高いが、高学年と中学では「品詞＋字種」の方が高い精度を示している。

3 つの文書特徴すべてを組み合わせた場合はほかのどの項目よりも高い精度を示しており、8 割から 9 割前後の精度が得られている。

表 4: 難易度判定の実験結果 (適合率/再現率)

文書特徴	低学年	高学年	中学
配当表	0.950 / 0.864	0.815 / 0.853	0.700 / 0.649
品詞	0.839 / 0.591	0.797 / 0.848	0.708 / 0.663
字種	0.844 / 0.614	0.746 / 0.898	0.733 / 0.483
字種＋品詞	0.909 / 0.682	0.821 / 0.898	0.787 / 0.683
配当表＋字種＋品詞	0.950 / 0.864	0.889 / 0.918	0.842 / 0.805

4 おわりに

本研究では、文書特徴を用いることによって教育コンテンツの難易度を自動的に判定する手法を検証した。漢字配当表の漢字の出現比率、品詞の出現比率、字種の出現比率を文書特徴とし、参考書データを用いた教育コンテンツの難易度判定の実験を行った結果、8 割以上の精度で低学年／高学年／中学の判定ができることを示した。

今後は、ホームページなど他のコンテンツを対象とした難易度判定実験や、語尾表現など他の文書特徴を用いた難易度判定手法の検討を行う予定である。

謝辞

本研究は通信・放送機構 (TAO) の直轄研究「学校インターネットにおける教育用情報検索技術の研究開発」の一環として実施しているものである。また、(株) 旺文社より小学参考書「達人シリーズ」、中学参考書「サンライズ」[3]の研究利用の許可を頂いた。関係各位の支援に感謝する。

参考文献

1) 文部省, 小学校学習指導要領, 平成 10 年文部省告示第 175 号 (1998).

2) C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery (1998).

3) 旺文社, 小学参考書「達人シリーズ」, 中学参考書「サンライズ」.