相対的観点に基づく類似難易度文書検索システムの構築

手塚智史　寺田博視　田中久美子
東京大学大学院情報理工学系研究科
{tezuka, terada}@cl.ci.i.u-tokyo.ac.jp, kumiko@i.u-tokyo.ac.jp

概要

語学教員や学習者にとって適切な難易度の文書を検索するシステムへの要求は大きい。

従来の難易度判定手法には、文書の難易度を表す回帰式を作るものや、文書をいくつかの難易度に分類する手法がある。一方、著者らの提案する方法では、機械学習を用いて 2 文書の相対的な難易判定を行う学習器を構成し、これを文書集合中の全 2 文書に適用し集合に順序構造を導入することで、文書の難易判定を行う。

本稿では、この相対難易判定を用いることで、web などから取得した大量の文書中から、ユーザの入力した文書と難易度上類似する文書を検索するシステムを構築したので報告する。

1　はじめに

語学教員や語学学習者にとって適切な難易度の文書を探すことは重要なことであり、多くの労力が払われている。そのため、適切な難易度の文書を検索するサービスやシステムへの要求は大きいものとなっている。

文書の難易度判定は長年研究されている分野であり、多くの手法が提案されてきている。筆者らは、相対的な観点から文書の難易判定を行う手法を提案した [1]。本手法を応用することで、web などから容易に取得することのできる多くの文書中から、難易度の観点で文書を獲得するシステムを構築した。

2　難易度判定

本システムでは、文書を難易度の観点から検索を行うために、文書の難易度判定手法を使用している。

2.1　従来手法

文書の難易度判定手法には Readability 研究と関連して過去に多くの研究がある [2–4]。

初期のものとして Flesch-Kincaid [5] や Dale-Chall [6] などがある。前者は単語や文の長さといった文書から得られる単純な統計に基づき難易度を判定する。後者は、このような統計に加え、基本単語リストを人手で構築し、ここに含まれない文書中の単語数といった統計を用いる。これらの手法は、回帰問題として文書難易度の判定を行っている。

最近の研究では、乾ら [7] による SVM [8] の回帰拡張である SVR を用いたものがある。この研究では 2 文書間の相対的な難易度を用いて SVR による機械学習を行い、構築した難易度判定器を用いて文書の難易度判定を行っている。また、文書の難易度判定を分類問題として扱ったものに、Collins-Thompson ら [9] による統計モデルを用いた手法や、Schwarm ら [10] による SVM を用いたものがある。

2.2　相対難易判定

筆者らは、相対的観点から文書の難易判定を行い、文書集合を整列させることにより難易判定を行う手法を提案した。

文書 a が文書 b よりも難しい場合 $a > b$ とする二項関係を考え、この二項関係を機械学習を用いて判定を行う。文書集合中の任意の 2 文書間の二項関係を判定していくことで、文書集合全体を難易度順に整列させることが可能である。具体的には、文書ペア a,b からベクトル V_{ab} を作成し、V_{ab} を入力したとき、$a > b$ の場合には $+1$ を出力し、$a < b$ の場合には　1 を出力する学習

器を構成する。この問題は 2 値分類問題であるので、学習器として SVM を使用する。

V_{ab} の構成には、まず文書 a,b から素性ベクトル V_a,V_b を作成し、ベクトルの何らかの演算 $\circ$ を行い、$V_{ab} = V_a \circ V_b$ とする。

素性と演算には様々な方法が考えられるが、本システムでは、素性は文書の単語の相対頻度と、それぞれの単語の大規模コーパス上での絶対頻度の対数を使用し、演算は素性ベクトル同士の連結を使用した。

3 類似文書検索システム Terrace

3.1 システム構成

本システムは、文書の検索モジュール、文書取得モジュール、文書データベースの 3 つの部分からなる。

文書検索モジュールは、ユーザからの文書の入力に対して、難易度の観点から類似する文書を検索対象文書の中から検索を行いユーザに返すものである。検索手法については§3.3 で後述する。

文書取得モジュールは、定期的に更新される web サイトより検索対象とする文書を取得するものである。

文書データベースは、検索対象とする文書を保存しておくデータベースである。検索対象の文書は§3.2 で後述する整列手法であらかじめ難易度順に整列された状態で保存されている。また、文書データベースは文書取得モジュールによって取得された文書を追加することで、日々更新される。

3.2 文書整列

本システムでは、文書の相対的難易判定を使用する。そのため、文書を検索可能にするために対象文書を整列させておくことが必要である。

本システムでは、対象文書の整列に挿入ソートの一種である二分挿入ソートを基本とする手法を用いた。手順としては、あらかじめ整列済みの集合に対して、新たに追加するものの挿入位置の検索を行い追加する。これを繰り返すことにより、集合全体を整列させることができる。二分挿入ソートでは挿入位置の検索に二分探索を用いており、少ない比較で検索が行える。

しかしながら、文書の比較精度は 100 %ではないため、少ない比較では正しい挿入位置の検索を行う事が出来ない可能性がある。そこで、比較の精度を上げるために、一回に近傍の複数の文書と比較を行い、それらを総合して判定を行う方法をとっている。これにより、比較の精度を保ち、正しい挿入位置の検索を行う事が出来るようにする。

また、この手法を用いることで、定期的に文書を追加する際も同様の手法で行うことができるため、本システムの運用上にも利点となる。

3.3 検索

本システムでは、相対的な観点から文書の難易判定を行っているため、検索により得たい文書と難易度が同程度の文書を入力として必要とする。

この入力文書を検索対象の文書集合中に挿入した際に、挿入候補となる位置の近傍の文書が難易度の類似する文書であり検索結果となる。

文書の検索には二分探索を基本とする手法を使用する。二分検索では比較回数が比較的少なく、本システムでは比較に時間がかかるため少ない比較であることは好ましい。

また、比較の精度を保つため、整列時と同様に一回に近傍の複数の文書と比較をし判定を行っている。

4 評価実験

本システムの評価として、文書集合の整列と、検索に関する 2 つの観点から評価を行った。実験は英語と日本語で行い、使用したデータは表 1 にまとめた。

学習データとして、英語では英字新聞の Time [11] と TimeForKids [12] を使用した (LD1)。日本語では朝日新聞の大人向け [13] と子供向け [14] のデータを使用した (LD2)。学習にはそれぞれのデータから無作為に 600 文書ずつ抽出し、LibSVM [15] を使用し機械学習を行った。

テストデータとして、英語では母国語向けとして Reading AtoZ [16] を使用し (TD1-M)、外国語向けとして、日本の英語教科書 [17,18] を使用した (TD1-F)。日本語では母国語向けとして国

表 1: 実験に用いる文書データ

英語データ			
ラベル	コーパス	レベル数	文書数
LD1	Time	2	600/600
TD1-M	Reading AtoZ	27	674
TD1-F	英語教科書	153	153
日本語データ			
ラベル	コーパス	レベル数	文書数
LD2	朝日新聞	2	600/600
LD2-M	国語教科書	58	58
LD2-F	日本語能力試験	4	44

語教科書 [19] を使用し (TD2-M)、外国語向けとして、日本語能力試験 [20] を使用した (TD2-F)。

テストデータはあらかじめ文書に難易度がつけられている。TD1-M はアメリカの小学校 1 年生から 5 年生の学年別に 5 段階に分かれており、それぞれの学年でさらに分けられ、合計 27 段階に分けられている。本稿ではそれぞれの段階をレベルと呼び、同じレベルの文書は難易度が同じとした。TD1-F と TD2-M は教科書の学習順に並んでおり、1 文書を 1 レベルとした。つまり、レベル数は文書数と同じになっている。TD2-F は 1 級から 4 級までの 4 段階に分けられている。D1-F は TD1-M よりも、TD2-M は TD2-F よりもレベル数が多くより難しい問題となっている。

実験では従来手法との比較として、Flesch-Kincaid、Dale-Chall、SVR を用いるものの 3 種類の手法 (§2.1) と比較を行った。ただし、Flesch-Kincaid と Dale-Chall は英語のみの手法なので TD1-M と TD1-F のみで行った。

4.1 整列実験

本システムで用いている二分挿入ソートを基本とした手法 (§3.2) による文書の整列精度の評価を行った。

評価方法として、テストデータを整列させ、スピアマンの順位相関係数を用いて評価を行った。

結果を図 1 に示す。横軸にはテストデータを、縦軸にはスピアマンの順位相関係数を示した。

TD1-F では、0.55 程度で Flesch-Kincaid より僅かによい結果であった。TD1-F を除くデータでは、提案手法により整列をさせたものが 0.8 以上の相関を示しており、ほかの手法に比べよい結果を得た。また、同じ手法で比べると Dale-Chall

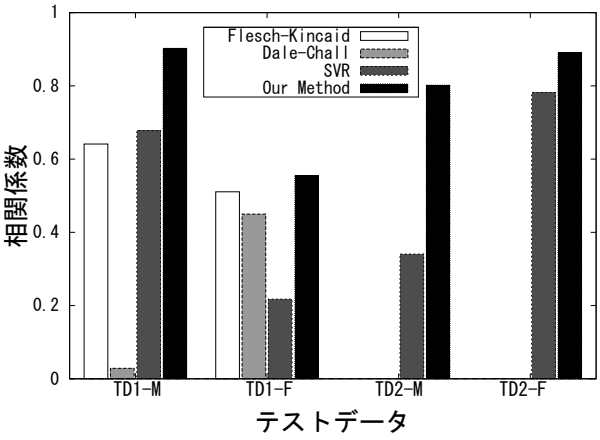

図 1: 整列結果

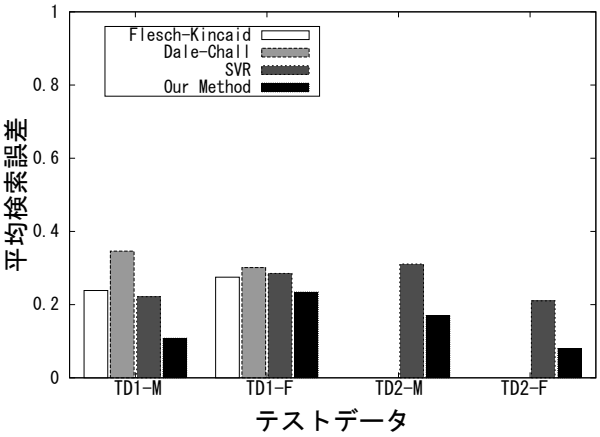

図 2: 検索結果

以外は、テストデータのレベル数が少ないほど相関が比較的大きくなる傾向がみられた。

4.2 検索実験

本システムで用いている、二分探索を基本とする手法 (§3.3) により難易度の類似する文書を検索する評価を行った。

評価方法として、テストデータから 1 文書を検索入力とし残りの文書を整列済みの検索対象として検索を行い、検索結果の文書と入力した文書のレベルの差をとった。全文書について同様の実験を行い、その平均をとり評価した。また、難易度の差はそれぞれのデータの全レベル数で割り正規化をしている。

結果を図 2 に示す。横軸にはテストデータを、縦軸に平均検索誤差として、正規化した入力文書と検索結果の文書のレベルの差の平均値を示した。

縦軸は誤差を表しているので、数値が小さい方がよりよい精度で検索が行えていることを表している。全てのテストデータに対し、本手法の方がほかの手法に比べ誤差が 40%〜90%ほどになっており、よい精度で検索を行うことができたという結果を得た。また、同じ手法を比べると Dale-Chall 以外は、テストデータのレベル数が少ないほど検索精度が比較的よくなるという結果を得た。

5 まとめと今後の展望

相対的観点からの文書難易判定を用いた文書検索システムを構築した。

提案手法による、類似難易文書の検索と文書集合の整列の精度を検証する実験を行い、従来手法を使用したものよりもよい結果を得た。

今後の展望として、本システムを Web サービスとして公開することを予定している。

また、現在は英語と日本語しか対応していないが、提案手法はどのような言語にも応用可能であるので、言語資源が取得できればほかの言語でも同様のシステムを構築することが可能である。今後は複数言語での実装を行っていきたい。

参考文献

[1] 寺田博視, 田中久美子. 文書の難易順序判定法. 言語処理学会第 14 回年次大会ワークショップ「教育・学習を支援する言語処理」論文集, pp.59-62, 2008.

[2] DuBay, W. (2004a). *The principles of readability*. Costa Mesa, CA: Impact Information.

[3] DuBay, W. (2004b). *Unlocking Language: The Classic Studies in Readability*. BookSurge Publishing.

[4] Klare, G. (1963). *The Measurement of Readability*. Iowa State University Press.

[5] Kincaid, J., Fishburne, R., and Rodgers, R. (1975). *Derivation of new readability formulas for Navy enlisted personnel*. Research Branch Report 8-75, U.S. Naval Air Station.

[6] Chall, J. and Dale, E. (1995). *Readability revisited: the new Dale-Chall readability formula*. Cambridge.

[7] Inui, K. and Yamamoto, S. (2001). "Corpus-based acquisition of sentence readability ranking models for deaf people." In *Natural Language Processing Paci c Rim Symposium*, pp. 205-212.

[8] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

[9] Collins-Thompson, K. and Callan, J. (2004). "A Language Modeling Approach to Predicting Reading Di culty." In HLT-NAACL, pp. 193-200.

[10] Schwarm, S. E. and Ostendorf, M. (2005). "Reading Level Assessment Using Support Vector Machines and Statistical Language Models." In Annual Meeting of the ACL, pp. 523-530.

[11] Time. **http://www.time.com**, accessed in 2008.

[12] TimeForKids. **http://www.timeforkids.com/**, accessed in 2008.

[13] 朝日新聞. **http://www.asahi.com/**, accessed in 2008.

[14] こどもアサヒ. 朝日小学生新聞. **http://www.asagaku.com/**, accessed in 2008.

[15] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM : a library for support vector machines", 2001.

[16] Reading AtoZ. **http://www.readingaz.com/**, accessed in 2008.

[17] 森住衛ほか (2007). *New Crown English Series 1, 2, 3*. 三省堂.

[18] 米山朝二ほか (2007). *Genius English CourseI, II, Reading*. 大修館.

[19] 宮地裕ほか (2008). 国語教科書. 光村図書. 小学 4 年〜中学 3 年.

[20] 日本国際教育支援協会 (2008). 日本語能力試験 **http://www.jees.or.jp/jlpt/en/**, accessed in 2008.