

クラウド向き大規模データ解析機構を用いた 教育・研究文献の難易度判定

靳 展[†] 田胡和哉[‡]

[†]東京工科大学大学院バイオ・情報メディア研究科

[‡]東京工科大学コンピュータサイエンス学部

1 はじめに

本研究では、テキスト解析分野における専門文献の難易度判定手法を提案する。本研究を通して、専門文献の難易度値を得るという目的である。文献の読み手にとって、文章の全内容を読まずに、難易度値とキーワードに従っていけば、この文章は読み手にとって、妥当かどうか分かる。そうすると、文章を読もうとする者に対して、文章を読むか読まないかという自主判定することができる。本研究に基づく難易度判定のメカニズムを設計し、難易度判定のシステムを構築した。

2 研究の目的

2.1 従来研究

二十世紀から始まる、テキストの難易度(リーダビリティ)に関する研究の中で、最も応用されたのは Rudolf Flesch が 1948 年に提案した一般成年の閲読テキスト用の Flesch 公式[1]であった。

近年では 2004 年に、中條ら[2]が、教育的視点に立った客観的な指標を用いて日英パラレルコーパスを構成している英語テキストおよび日本語テキストの難易度を計測することによって、難易度によるテキスト分類の手法を提案している。

この分野における最新の研究は佐藤ら[3]が 2008 年から研究している日本語テキストの難易度の自動推定である。この研究では、日本語テキストの読みやすさ(やさしさ、難しさ)を自動的に柔軟に判定する手法が提案されている。現在、13 段階の難易度判定は実現されており、実社会の応用が試みられている。

2.2 専門文章の解析

これまで文章の易読性、あるいは言語自体の難易度推定について、研究が行われている。過去の

推定手法により、一定な計算を行った後に、日本語および英語文献の読みやすさの結果を得ることができるようになっている。

しかし、それらの研究では、日本語の学習者を対象として難易度を推定することを想定している。専門用語を含んでいる場合では、専門用語の難易度推定に関わる研究は行われていない。

例えば、一人の大学学部生に対象として、Java 基礎知識の入門文章と J2EE 技術の解説文章の難易度は全く違うものである。もし、ある程度の推定手法を生かして、学生にとってこの二つの文章の理解さ、要するに文章の難易度の順番を提供することが可能となれば、この学生にとって、確実に有効な参考文献となる。同じく、学生だけではなく、専門文献の難易度推定手法を通して、さらに、大学教育や研究事業などにおいても重要な技術支援を行うことができる。

3 提案手法

上記の専門用語を含まれる文献を解析するため、本稿は専門文献の難易度を判定する手法を提案する。

3.1 難易度判定

難易度判定のメカニズムの構成を図 1 に示す。

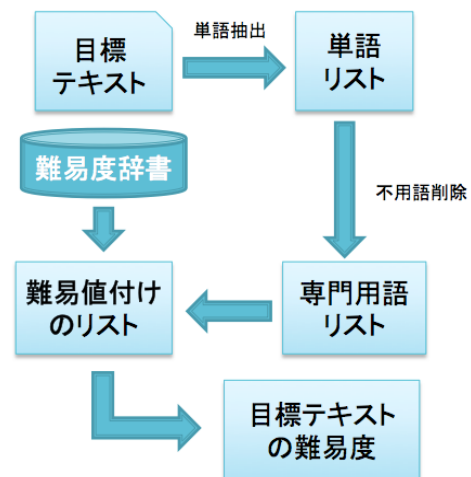


図 1 難易度判定メカニズムの構成図

The difficulty judging of the education and research literature using the large-scale data-analysis framework

[†] Zhan Jin (kinten0902@gmail.com)

[‡] Kazuya Tago (ktago@cs.teu.ac.jp)

School of Computer Science, Tokyo University of Technology

([†])

1404-1 Katakura, Hachioji, Tokyo 192-0982, Japan

本研究の判定手法は、単語分割手法と TF-IDF 法の技術を利用し、目標を達成する。具体的な手法は以下4つのステップに分けられる。

1. まず目標文献の中の単語を抽出し、各単語の出現頻度のリストを作成する。
2. そして、ノイズ除去処理を行い、専門用語以外の単語を削除する。
3. その後、事前に用意された難易度辞書を利用し、リストの中の各単語に難易値を付ける。
4. 最後に、難易値付けのリストを統一するとともに、唯一の文献難易度を判定する。

3.2 難易度辞書の作成

難易度辞書の有効性は最後の結果に関わる一番重要な要因だと考えられる。そのために、なるべく有効性が高い難易度辞書を生成した方がよい。難易度辞書の生成については既に複数の方法が提案されているが、本研究は以下の二種類方式で生成した難易度辞書を選択し、最後の結果を比較して認証する。

1. 個人の知識範囲によって、人工的にカスタム難易度辞書を定義する。この方式を利用し、最後の難易度判定結果の精度が高いと考えられる。しかし、辞書の作成時間が長く、効率は低い。
2. ある文献を比較標準として選択し、文献内容の専門用語に対する、難易度辞書を生成する。この方式は構築されたシステムが自動的に難易度辞書を生成するので、方式1より効率が良いと考えられる。

4 設計と実装

本システムの構成を図2に示す。

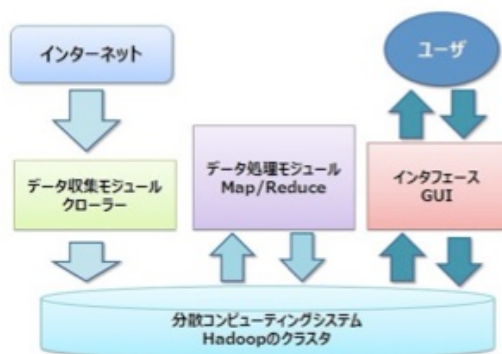


図2 システムの構成図

本システムは大きく分けて4つの部分から構成されている。

1. Apache が提供した分散コンピューティングフレームワーク Hadoop[4]を基づいたクラウド基盤を構築する。

2. データ収集モジュールは HTMLParser と HttpClient この2つライブラリを導入し、文献に特化したクローラを実装した。指定した URL から適当な文献ファイルをダウンロードし、Hadoop の HDFS に保存する。

3. 分析モジュールは難易度判定やキーワードなどいくつかのテキスト解析機能を実装している。それらを用いることで、分析モジュールを経由して HDFS に保存されたデータに対し、短時間内で各自の難易度を算出できる。

4. インタフェースは RPC フレームワーク Thrift を利用し、実行結果をシステム使用者に示す。

本研究では、12 台のサーバに Hadoop をインストールし、クラスターを構築した。システム収集モジュールの文献に特化したクローラをローカルで実行し、初回実験用のデータが集まった。二種類の難易度辞書を提案し、難易度推定のメカニズムのコーディングが完了した。

現在、試行を行っている段階である。第一種類の難易度辞書の生成方式で Java 言語キーワードの難易度辞書を作成は「1:かなり易しい, 2:やや易しい, 3:普通, 4:やや難しい, 5:かなり難しい」という5段階に分け、サンプル文章は Java 技術に関連する専門文献を五冊選び、難易度推定のメカニズムを利用し各文献の自難易度の結果が生成された。

5 おわりに

今後、大規模データ処理に対する信頼性と難易度判定結果の有効性を検証する予定である。また、本システムに対し、改善する必要があるものはいくつか存在する。例えばマシンラーニングの技術を加えれば、難易度推定の精度が上がると考えている。

参考文献

- [1] Flesch R. A new readability yardstick. Journal of Applied Psychology 32, pp:221-233 (1948)
- [2] 中條清美, 白井篤義, 内山将夫. 日英パラレルコーパスを構成するテキストの難易度分類に関する研究. 日本大学生産工学部研究報告 B, pp:57-68 (2004)
- [3] 近藤陽介, 松吉俊, 佐藤理史. 教科書コーパスを用いた日本語テキストの難易度推定. 言語処理学会第14回年次大会発表論文集, pp:1113-1116 (2008)
- [4] <http://hadoop.apache.org/>