

# Wikipedia における編集履歴を利用した編集者の能力推定

安井 貴規†

藤田 桂英‡

† 東京農工大学 工学部

‡ 東京農工大学大学院 工学研究院

## 1 はじめに

Wikipedia は、世界最大の集合知を形成する Web サービスである。Wiki と呼ばれるオンライン上で記事が編集できるシステムにより誰でも手軽に編集できる一方で、すべての編集者が Wikipedia の方針や記事自体に関する知識を十分に持っているとは言えない問題がある。また、Wikipedia 自体に編集者のモチベーションを向上させる直接的なシステムも存在していない。

本論文では、Wikipedia の編集者が追加した文章および記事のメタ情報を利用して、良質な Wikipedia の記事を形成するユーザとしての能力を自動推定するランダムフォレストによる分類モデルを提案する。また、分類器の分類結果から得られる F 値や、モデルの特徴素の重要度から、本モデルの評価を行う。

## 2 Wikipedia における編集者の能力推定手法

### 2.1 リビジョンのフィルタリング

特徴素をテキストやメタデータから抽出する前にリビジョンのフィルタリングを行い、学習や分類に際し不要なリビジョンを除外する。

まず、初版から最新版に至るまでの間に取り消しが行われた、または取り消しを行ったリビジョンを除外する。取り消しに関するリビジョンは最新版では、編集が無かったことにされているため除外する。Wikipedia が採用したソフトウェアである MediaWiki には、特定のリビジョンを取り消し、記事を対象となるリビジョンの直前の状態に戻す機能がある。リビジョンの取り消しの検出は、リビジョンに紐付けられたテキスト ID の値が直前のリビジョンの値より小さくなったことを検出することにより可能である。

次に、人間が直接記事を編集しない Bot アカウントや、ログインを行わずに匿名として記事の編集を行ったことを示す IP ユーザによる編集を除外する。IP ユーザの除外を行うのは、一人の人間による編集であることが保証できないためである。Bot アカウントの抽出には、「Wikipedia:Bot/ステータス」という Bot アカウントの一覧が記載されている記事の内容をスクレイピング

した。IP ユーザの除外は、データベース上のユーザ ID が 0 で統一されているために判断できる。

図 1 に、4 つのテキスト ID と 5 つのリビジョンにより構成される「正史」(リビジョンのフィルタリングによって除外されなかったリビジョンの集合)の例を示す。例では、取り消しが行われたリビジョン  $r_2$  と、 $r_2$  の取り消しを行ったリビジョン  $r_3$  が「正史」から除外されている。取り消しの操作は、 $r_3$  にてテキスト ID が  $t_3 (> t_2)$  となるべきところ  $t_1 (< t_2)$  となっていることから検出できる。

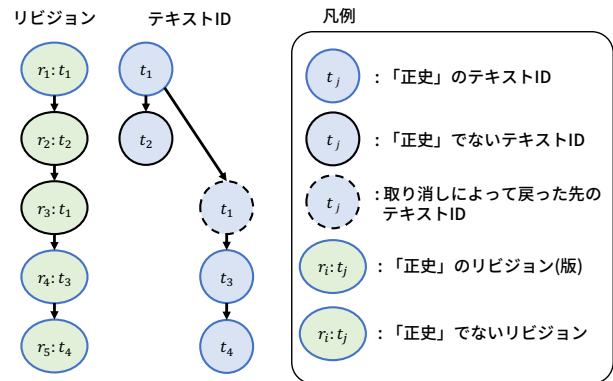


図 1: 4 つのテキスト ID と 5 つのリビジョンで構成される「正史」の例

### 2.2 IEF 特徴量

IDF (Inverse Document Frequency) 特徴量は、複数文書間における単語の重要度として知られている。そこで本研究では、IDF 特徴量を文書単位ではなく編集者単位で扱う IEF (Inverse Editor Frequency) を特徴量として定義する。Daxenberger ら [1] の研究では本文やメタデータを用いた編集内容の分類を行い、本文に関する特徴が分類に大きく影響したことから、本研究でも有用な特徴量となることが期待される。

IEF は単語  $w$  とそれぞれの編集者が追加した文章集合の集合  $S$  より、式 (1) によって求められる。

$$IEF(w, S) = \log \frac{|S|}{\{|s : w \in s\}} \quad (1)$$

### 2.3 編集行数比率

編集の方針や記事内容自体に詳しい編集者の編集は最新版まで残存することが期待されるから、ある編集

An Estimation of Editor Skills using Edit History in Wikipedia  
 †Faculty of Engineering, Tokyo University of Agriculture and Technology  
 ‡Institute of Engineering, Tokyo University of Agriculture and Technology

者が追加した行数を特徴素として取り入れた。編集行数比率を算出する前に、各行がどの編集者によって編集されたものであるのかを解析する。この解析は、初版から最新版に至るまでに出現したリビジョンおよび紐付けられたテキストの差分を取ることで実現できる。ただし、取り消しを行ったリビジョンや取り消しが行われたリビジョンは除外する。

### 3 評価実験

#### 3.1 日本語版 Wikipedia からの正解データの作成

2018年7月1日に作成された日本語版 Wikipedia のダンプファイルを用いて、記事と編集者のセットに対する能力ラベルを9人の参加者により1,2,...,6の6段階 (Higher is better) で作成した。記事は「秀逸な記事」から30記事 (編集者との組み合わせは1,239通り) を選択し、10記事ごとの3グループに振り分けた。それぞれのグループに対し3人ずつを割り当て、中央値を代表値として利用した。表1に示すように、1,2,...,6の6段階から「1」、「2」、「3または4」、「5または6」の4段階へ再割当てを行った後のアノテーションデータ全体の Fleiss' Kappa 一致率は0.51となった。

表1: 4段階評価でのアノテーションの  $\kappa$  一致率

グループ	$\kappa$	Landis ら [2] の解釈
0	0.41	並の一致
1	0.51	並の一致
2	0.63	相当な一致
全体	0.51	並の一致

#### 3.2 実験結果

表2に本論文の提案モデル (Proposal) の学習結果から得られる、各特徴素の重要度を示す。表2から、合計 IEF 値が最も分類に寄与していることがわかる。一方でリビジョン取り消しに関わる2個の特徴素の重要度はそれぞれ0.04, 0.01となった。リビジョンの取り消しに関わる特徴素の重要度が低くなった原因として、リビジョンの取り消し自体があまり行われていなかったことが挙げられる。

提案モデル (Proposal) との比較のためのベースラインとして、データセットの能力ラベルの分布と同一の分布となるようにランダムに推測するモデル (Baseline) を設定した。また、提案モデルのほかに、提案モデルから合計 IEF 値の特徴量のみを除いたモデル (Proposal without IEF) も比較対象として設定し、IEF 特徴量の有用性を確認した。

表2: 学習済みモデルから得た各特徴素の重要度

特徴素	重要度
合計 IEF 値	0.48
編集回数比率	0.22
編集行数比率	0.14
記事内編集者数	0.11
リビジョン取り消し回数	0.04
リビジョン取り消され回数	0.01

表3にそれぞれの手法による分類結果のF値を示す。表3から、提案モデルの結果はベースラインの結果と比較してF値が0.217 (70.2%) 向上したことが確認できる。また、合計 IEF 値を除いた手法と比較して、提案モデルのF値は0.074 (16.3%) 高かった。

表3: それぞれの手法によるF値

手法	F 値
提案手法	<b>0.526</b>
IEF を除いた提案手法	0.452
Baseline	0.309

### 4 まとめ

本論文では、テキストやメタデータから得られる各種特徴素を利用して、Wikipediaの記事における編集者の自動能力推定を行うモデルを提案した。また、評価実験では、データセットの分布に従ってランダムに分類を行うベースラインと比較を行い、提案モデルが有用であることを確認した。さらに、合計 IEF 値を特徴素から除外したモデルと比較した結果、テキストから得られる特徴素がメタデータから得られる特徴素より効果的に作用していることを確認した。

メタデータから得られる特徴素で有効なものは少なかったため、特徴素の変更や追加など、改善の余地が残されている。今後はWikipediaに限らず、複数の投稿者や編集者が存在するサービスにおける能力推定について注目していきたい。

#### 参考文献

- [1] Johannes Daxenberger and Iryna Gurevych. Automatically classifying edit categories in Wikipedia revisions. In *EMNLP*, 2013.
- [2] J Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–74, 04 1977.