7S-08

回帰分析によるオンライン小説の人気度推定

Popularity estimation of online novels using regression analysis

実崎 直人[1)]　**伊東 栄典**[2)]

Naoto Jitsuzaki　Eisuke Ito

1　はじめに

近年，ネットで動画・音楽・小説などのコンテンツを自由に投稿・公開できるようになった。これらの利用者がコンテンツを投稿するサービスは，CGM (Consumer Generated Media) と呼ばれる。CGM サイトには，YouTube やニコニコ動画，小説家になろう，comico などが存在する。CGM サイトには毎日多数のコンテンツが投稿されており，膨大な利用者が閲覧して人気である。

我々は CGM の「ニコニコ動画」を対象に，利用者コメントに基づく動画推薦，動画再生回数の推定 [1] など研究してきた。また「小説家になろう」の小説を対象に，読者と小説のリンク構造に基づく小説推薦，小説キーワードの分散度調査 [2]，偶発性を重要視する小説推薦 [3] を研究してきた。

ネット上の CGM 小説の多くは低品質であるものの，ごく一部に高人気かつ高品質な小説が有る。高人気小説は，印刷物として出版・販売されたり，さらには漫画やアニメに展開されるものも有る。将来人気になる小説を発見できれば，個人への小説推薦や書籍・漫画業界での展開にも役立つ。本研究では，将来人気になる作品を発見を目指すため，「小説家になろう」の小説群を対象に，回帰分析による小説の人気度を推定を行う。本論文では，人気度の定義，回帰分析手法，用いた説明変数，および推定結果について報告する。

2　小説家になろう

「小説家になろう (http://syosetu.com/)」は，株式会社ヒナプロジェクトが提供する小説投稿サイトである。誰でも小説閲覧可能であるものの，利用者登録により小説投稿や，小説ブックマーク，作者および小説へのコメント投稿が可能になる。2004 年の開設当初は個人サイトであったが，アクセス増により 2008 年からグループ運営に移行し，2010 年に正式に法人化した。Wikipedia [4] によると，2014 年 12 月時点のアクセス数は月間約 9 億 5000 万 PV，ユニークユーザー数は 400 万人である。また 2018 年 12 月 27 日，登録者数が 1,431,306 人，掲載小説数は 618,761 作品である。なおサイトの小説は「なろう小説」と呼ばれる事が多い。

2.1　小説メタデータ収集

利用者が作品を閲覧する際，小説の題名・作者・あらすじなどを参照する。小説を説明するデータを「メタデータ」と言う。「小説家になろう」のでは，題名，作者名，あらすじ，レビュー数，キーワード，また人気尺度であるブックマーク数や総合評価点が含まれる。本研究では，回帰分析にメタデータに含まれる数値や単語数を用いる。

小説メタデータの収集には「なろう API」を用いた。「なろう API」は，「小説家になろう」を運営しているヒナプロジェクト社が提供する REST 型の Web API である。Python 言語でメタデータ収集クローラーを作成し，全小説のメタデータを集めた。なろう API では，小説メタデータの形式として JSON か YAML が選択できる。今回は JSON 形式で取得した。2018 年 11 月 9 日までに収集した小説数は 521,095 件である。これを分析に用いる。

2.2　Elastic Stack によるデータ管理と分析

Elastic 社はデータ管理分析のためのオープンソース製品群である「Elastic Stack」を提供している。「Elastic Stack」には 4 つのソフトウェア（Elasticsearch，Kibana，Beats，Logstash）が含まれている。本研究では，収集した小説メタデータの保存・検索・分析のために Elasticsearch を用い，データの可視化に Kibana 用いた。なお，自作の Python プログラムで JSON 形式の小説メタデータを Elasticsearch に投入した。

3　回帰分析による人気度推定

3.1　目的変数と説明変数

人気度の推定について述べる。「なろう小説」の人気度には，メタデータに含まれるブックマーク数と総合評価点の 2 つが利用できる。2 つの相関は 0.9 以上で高い。今回はブックマーク数を人気度とした。回帰分析の目的変数は，初投稿日から最終投稿日までの 1 日毎の平均ブックマーク増加数の対数値とした。ブックマーク数の分布は対数正規分布になるため [2]，対数値で正規分布になるようにした。

メタデータに含むもので回帰分析の説明変数となりうる値は，(a) 会話率，(b) レビュー数，(c) 小説文字数，(d) 読了時間，(e) ユーザ ID（作者 ID），(f) 全掲載部数，が有る。部数（話数）が多いものや文字数が多いものを好む読者も多い。また作者で小説を選ぶことも多いため，説明変数とした。更にメタデータに含まれない 2 つの値，(g) 初投稿日からの取得日までの経過日数と，(h) 初投稿日から最終投稿日まで 1 日あたりの掲載部数，を算出して説明変数に加えた。連載頻度の高いものが好まれるため (h) を含めた。また連載中断状態の小説が多いため，(g) と (h) で中断を表現できると考えた。

3.2　推定精度の指標

推定精度の指標には回帰分析の決定係数 R^2 と，平方平均二乗誤差率 RMSPE (Root Mean Squared Percentage Error) を用いた。式 1 に決定係数を示す。式 1 で，y_i は各実測値，f_i は回帰分析による推定値，μ は実測値の平均値である。

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \mu)^2} \tag{1}$$

式 2 に平方平均二乗誤差率 RMSPE (Root Mean Squared Percentage Error) を示す。

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - f_i}{y_i + 1}\right)^2} \times 100 \tag{2}$$

1)　九州大学工学部電気情報工学科
2)　九州大学情報基盤研究開発センター

3.3 線形回帰分析による人気度推定

まず，短編を除いた小説をジャンル毎，次にジャンル毎のブックマーク数上位 10%のそれぞれに線形重回帰分析として，リッジ回帰を適用した。重回帰分析とは，2 つ以上の説明変数から 1 つの目的変数を推定するための回帰分析法である。

表 1 にジャンル毎，表 2 にジャンル毎のブックマーク数上位 10%の決定係数と平方平均二乗誤差率を示す。上位 10%を取っても，決定係数に改善は見られず，推定できていない。平方平均二乗誤差率 (RMSPE) については，一部で改善がみられたが，推定精度が上がったとは言えない。

表 1 ジャンル毎の線形回帰分析

ジャンル	決定係数	RMSPE
異世界	0.094	70.8
現実世界	0.050	46.1
ハイファンタジー	0.118	36.6
ローファンタジー	0.097	28.3
ヒューマンドラマ	0.055	28.2
コメディー	0.005	35.8
ホラー	0.125	39.5
純文学	-0.014	35.2
アクション	0.148	22.5
推理	0.086	37.2
歴史	0.083	43.6
空想科学	0.167	29.8
VR ゲーム	0.062	53.8
パニック	0.099	32.9
宇宙	0.052	33.2

表 2 ジャンル毎の上位 10%の線形回帰分析

ジャンル	決定係数	RMSPE
異世界	0.003	38.2
現実世界	0.029	32.9
ハイファンタジー	0.040	51.8
ローファンタジー	0.021	35.7
ヒューマンドラマ	0.001	21.7
コメディー	0.017	30.0
ホラー	0.011	12.0
純文学	-0.006	11.8
アクション	0.033	26.7
推理	-0.120	18.0
歴史	-0.025	40.1
空想科学	0.042	22.1
VR ゲーム	-0.018	33.0
パニック	-0.031	26.7
宇宙	-0.010	31.7

3.4 非線形回帰分析による人気度推定

次に，サポートベクター回帰 (SVR，Support Vector Regression) の RBF カーネルによる非線形重回帰分析を適用した。使用データは線形回帰分析でのものと同じである。

表 3 に結果を示す。非線形回帰分析では，正確な推定はできていないが全体的に推定精度に改善がみられた。特に，異世界と VR ゲームにおいては大きく改善した。図 1 は，決定係数の最も高いジャンルの異世界について，縦軸を予測値，横軸を実測値として描画したものである。

表 3 ジャンル毎の上位 10%の非線形回帰分析

ジャンル	決定係数	RMSPE
異世界	0.613	20.4
現実世界	0.329	24.1
ハイファンタジー	0.292	31.9
ローファンタジー	0.116	25.3
ヒューマンドラマ	0.115	18.7
コメディー	0.208	22.5
ホラー	0.278	15.5
純文学	0.218	12.3
アクション	0.098	30.1
推理	0.083	13.3
歴史	0.266	33.9
空想科学	0.224	19.8
VR ゲーム	0.543	21.1
パニック	0.118	26.3
宇宙	0.122	31.5

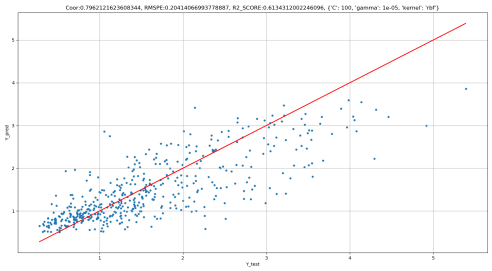

図 1 「異世界」小説の実測値と予測値の関係

4 おわりに

本研究では「小説家になろう」の小説群を対象に，回帰分析による小説の人気度を推定した。目的変数となる人気度としてブックマーク数の対数値を用いた。

単純な線形回帰分析と，RBF カーネルによる非線形重回帰分析を適用した。線形回帰分析では，決定係数も誤差率も悪く，人気度を推定できていない。RBF カーネルによる非線形重回帰分析を適用した所，決定係数および誤差率の値は改善したものの，推定不十分である。

今後は題名，あらすじ，最初の数話における単語出現頻度の利用も検討したい。小説メタデータは定期的に収集しているため，時系列データの回帰分析のように，数年前，1 年前，数ヶ月前，などの過去の値を，現在の人気度推定に用いることも行いたい。読者感想の盛り上がりも人気に影響する可能性があるので，感想の利用も検討したい。

参考文献

[1] 柴田知親，伊東栄典：回帰分析による CGM 動画再生回数推定，第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM2018)，pp. C5–2 (2018).

[2] Ito, E. and Honda, Y.: Keyword diversity trend of consumer generated novels, *Proceedings of ICESS2017* (2017).

[3] 飯田委哉，伊東栄典：セレンディピティを考慮した CGM 小説推薦，人工知能学会第 15 回データ指向構成マイニングとシミュレーション研究会，pp. 2–0284 (2018).

[4] Wikipedia: 小説家になろう in Wikipedia，https://ja.wikipedia.org/wiki/%E5%B0%8F%E8%AA%AC%E5%AE%B6%E3%81%AB%E3%81%AA%E3%82%8D%E3%81%86.