Wikipedia ページ閲覧回数に基づく知名度推定

小林　周平†　　田島　敬史††

† 京都大学情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
†† 京都大学情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †skobayashi@dl.soc.i.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　YouTube 等のソーシャルメディアを通じて多くの人が自身のコンテンツを発信するようになり，自身の知名度を知りたいというニーズは高まっていると考えられる．しかし, 知名度はアンケート等によって調査する以外に現状は方法がなく，またアンケートの対象となるような人も限られてしまうため，そのニーズを満たすことはできない．そこで，インターネットユーザの Wikipedia ページ閲覧回数に基づいて知名度を推定する手法を提案する．閲覧行動には話題性と定常性という二つの性質があると仮定し，それらの性質を用いることでユーザの知名度の推定を行う．

キーワード　Web 情報, ソーシャルメディア, 知名度, ユーザ検索モデル,Wikipedia

1　はじめに

Youtube 等のソーシャルメディアの普及で，自身の作成したコンテンツを発信する人が増えている．また，その自身が発信したコンテンツをもとに収益を獲得し，生計を立てているユーザも増えてきている．収益を増やすうえで大切なこととして，多くの人に自身のコンテンツを閲覧してもらう必要があり，したがって自身がどの程度認知されているのか，知名度を知りたいと考えているコンテンツ発信者は多いと考えられる．しかし現状，知名度を調査するにはアンケートを実施する必要があり，それにはコストがかかってしまう．また，知名度調査は企業などによって行われてはいるが，その調査対象となるような人は限られてしまっている．そのため，自身の知名度を一個人が調べることは現状困難であり，現実的ではない．

そこで，知名度をインターネット上に存在する情報のみから推定することができれば，知名度調査の対象とならないような人も自身の現状の知名度を知ることができるようになる．さらには，現状の知名度推定にかかるアンケート等のコストも削減することができる．

そのため，本研究ではインターネットユーザの検索行動を用いて知名度を推定できないかを検証した．知名度推定の方法として，インターネット検索結果数が挙げられると Wikipedia の知名度のページに記載されているが，それは誤解であり，小紫と田島 [1] が行った研究において，検索結果数と知名度には相関があまりないということが示されている．

本研究では，ある人物がインターネット上で検索されるのには二つのパターンがあると仮定し，その二パターンに検索行動を分割することによって知名度の推定を行った．検索行動には話題性と定常性の二つのパターンがあると仮説を立てた．話題性とは，テレビなどのメディアでその人物が話題に上ったため，インターネット検索される回数が増える状態のことを指す．例えば図 1 のように，俳優の Irrfan Khan の Wikipedia 英語ページの閲覧回数は，一日あたりの平均は 6,394 回である一方で，

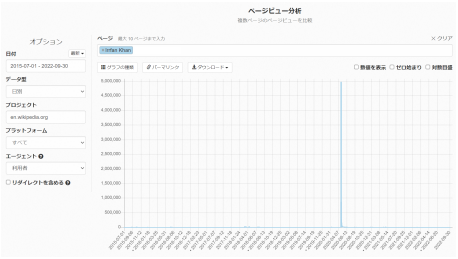

図 1　**Irrfan Khan の Wikipedia ページ閲覧数．**2015 年 7 月 1 日から 2022 年 9 月 30 日の期間において，Irrfan Khan が亡くなった 2020 年 4 月 29 日の閲覧数が非常に大きくなっている．(Irrfan Khan は本研究の実験では使用していない.)

2020 年 4 月 29 日には 4,968,949 回となっている．これは，この日に Irrfan Khan が亡くなっているため，そのことが影響していると考えられる．このように，テレビなどのメディアで話題になった際には人々は検索し，その結果閲覧数が多くなると考えられる．逆に，閲覧数が多い場合には世間で話題となっており，多くの人の目に触れたと考えることができる．したがって，話題性による閲覧数は，その人物についての知名度を向上させることに寄与している可能性があると考えられる．

定常性は，世間で話題に上っていない状態のことを指す．定常性による検索は，その人物に対して調べたいことがある個人によって行われるものであり，ある程度の検索数の幅にとどまっていると考えられる．ただ，知名度が高い人ほど検索しようとする母数が多くなるため，知名度が高い人ほど定常性による検索数の平均は高くなると考えられる．これら二つの性質に検索行動を分離し，知名度推定を行った．

本研究では検索数の代わりに Wikipedia の閲覧回数のデータを用いた．これは，検索数を直接取得することが困難であり，Wikipedia の閲覧回数は検索数と高い類似を持つ可能性が高いと考えられるからである [2]．また，YouTuber や TikToker 等，自身の知名度を知りたいと考えるような，ある程度活動が認知されているユーザは Wikipedia ページが作成されている可能性が高い．そこで，Wikipedia の閲覧回数を定常性と話題性に分

割して入力データとし，正解データとして知名度を用いて線形回帰を行った．

より詳しくは，過去の期間を複数の区間に分割し，各区間の閲覧数を一つの特徴量とすることで線形回帰の入力となる特徴ベクトルを作成した．その際，様々な期間の分割の仕方について実験を行った．また，定常性による閲覧数と話題性による閲覧数の情報をどのように組み合わせるかについても，四つの組み合わせ方について実験を行った．さらに，閲覧数の特徴量への変換方法についても三つの方法を比較した．また，正解データとして様々な世代ごとの複数の知名度を用いた．

その結果，定常性と話題性に分割し，同時に利用することによって全世代に対する知名度推定の精度が向上すること，対象とする知名度の世代が若くなるほど閲覧数の情報を用いた知名度推定の精度が上昇するということが明らかになった一方，閲覧数の情報のみで知名度を高い精度で推定することは，少なくとも単純な手法では困難であるということが判明した．

本論文の以降の構成としては，2 章において関連研究を，3 章において話題性と定常性について述べる．その後，4 章で提案手法の詳細，5 章で実験について述べ，6 章で本研究のまとめを行う．

2 関 連 研 究

本章では関連する先行研究について記載する．

2.1 カテゴリごとの偏りを考慮した検索結果からの知名度推定

有名人の知名度を推定する研究として，小紫と田島 [1] の研究がある．この研究では，有名人の性質として Web 親和性，話題性，蓄積時間と検索結果数を用いて知名度推定を行うモデルが提案されている．本研究は Web ページの検索結果数ではなく閲覧回数の情報を用いているため，知名度推定へのアプローチ方法が異なっている．検索結果数は Web 上の情報の量を表し，情報の供給量を反映しているのに対し，検索回数や Wikipedia の閲覧数は情報の需要の多さを反映している．Web 上でのある情報に関する供給と需要は必ずしも一致するとは限らず [3]，一般の人々の間での知名度推定のためには，情報の需要の多さの方がより有用なのではないかと考えられる．

2.2 検索結果のメタデータ分析に基づく人物ランキング

検索エンジンを用いて人物のランキングを行う研究として，Ma と Yoshikawa [4] の研究が挙げられる．この研究では，与えられた人々のランキングを行う際，各人物の知名度の高さのスコアを用いてランキングを行っている．知名度の高さのスコアを求める際には，URL，スニペット，検索結果数といった検索結果のメタデータを用いている．知名度のスコアを求めるという点は本研究と類似しているが，本研究では知名度を求める際に Web ページの閲覧回数の情報を用いているという点，また知名度をそのまま求めようとしている点で異なっている．

2.3 検索頻度推定のための Wikipedia ページビューデータの分析

Wikipedia のページビューと検索頻度の類似性について，吉田ら [2] が研究を行っている．この研究では，Google Trends の情報と Wikipedia ページビューの情報を用いて，Wikipedia ページへのアクセス数と検索頻度についての調査を行っている．調査の結果，Wikipedia におけるアクセス回数が多いキーワードのページの閲覧数と，そのキーワードの検索頻度には高い類似が確認された．本研究ではこの調査結果を利用し，取得が困難な検索回数の代わりに Wikipedia のページ閲覧数を用いている．

2.4 インターネット上の情報の実世界への応用

ソーシャルメディア上に散らばる情報など，インターネット上から得られる情報を用いて実世界の情報を推測する研究は広く行われている [5] [6]．

例えば，Tuarob ら [7] は感染症のダイナミクスをモデル化するために使用される従来の手法に，ソーシャルメディアから取得した情報を取り入れることにより，従来のモデルよりも高い精度で現実のインフルエンザ感染率を追従できることを示した．Sakaki ら [8] は Twitter 投稿をもとに，地震発生の検知や台風の軌道の推定などを行っている．その他，Kryvasheyeu ら [9] は，ハリケーン「サンディ」の進路への近接度とハリケーン関連のソーシャルメディア活動の間に強い関係があることを発見し，現実の脅威が，Twitter のメッセージストリームの強度と構成を通じて直接観察可能であることなどを示した．

3 話題性と定常性

この章では，定常性と話題性についての説明と，それら二つと知名度の関係について述べる．

3.1 知名度の正解データ

今回の研究では，知名度の正解データとして YouGovAmerica [10] が提供している，The Most Famous All-time Actors (Q4 2022) から抽出した 61 人の俳優の Fame を正解知名度とした．61 人は 10 位，20 位，30 位，…と 10 位間隔で知名度の順に選択し，そのうちアメリカの俳優のみを取得した．ここでアメリカの俳優のみを選択した理由としては，アメリカ以外の俳優の場合，アメリカ人以外の英語話者の閲覧数の影響が強くなる可能性が高く，その場合アメリカ人に対する知名度を正解として扱っている本研究では不適切であると考えたからである．Fame は 4 種類使用し，それぞれ All, Millennials, Gen X, Baby Boomers となっている．Millennials は 1982-1999 生まれ，Gen X は 1965-1981 生まれ，Baby Boomers は 1946-1964 生まれの世代に対する知名度を表す．

3.2 話題性と定常性の定義

本研究では，話題性とは Wikipedia の閲覧回数が通常時より多い状態を指し，定常性とは通常時の閲覧数を指すものとする．Wikipedia のページ閲覧回数は，Wikipedia のツールであ

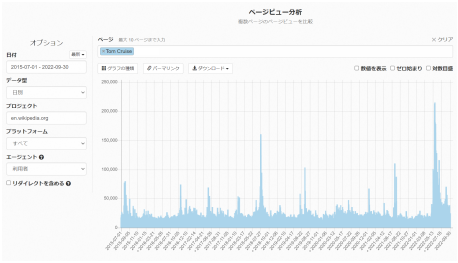

図 2　**Tom Cruise の Wikipedia ページ閲覧数．**オプションは，期間は 2015-07-01 から 2022-09-30 まで，データは日別，プロジェクトは en.wikipedia.org, プラットフォームはすべて，エージェントは利用者とした．同様のオプションにおける閲覧データを 61 人の俳優からも取得した．(Tom Cruise は実験では使用していない．)

る Pageviews Analysis を用いて取得した．取得する際，オプションで期間は「2015 年 7 月 1 日から 2022 年 9 月 30 日まで」，データ型は「日別」，プロジェクトは「en.wikipedia.org」, プラットフォームは「すべて」，エージェントは「利用者」とした．プロジェクトとは対象とするページの言語を，プラットフォームとはデスクトップ，モバイルアプリ，モバイルウェブといった閲覧方法を，エージェントとは対象とする閲覧者を意味している．

図 2 に見られるように，閲覧数はある一定付近の値をとる場合と，突然大きい値をとる場合があり，この一定付近の値をとっている状態を定常性と呼び，急激に大きな値をとる状態を話題性と呼んでいる．話題性の原因としては，俳優に実世界で何らかの出来事が発生したことが考えられ，例えば俳優が死亡した際や，俳優が出演した映画が公開されたなど，テレビ等メディアでの露出が増えた際に，その俳優についてインターネット検索が行われるために起こると考えられる．今研究では定常性と話題性に閲覧数を分離して実験を行うため，それぞれの俳優の閲覧数に対して以下のような式を用いて分離を行った．

$$\begin{cases} \text{閲覧数} \geqq \text{第三四分位数} + \text{四分位範囲} \times 1.5 & \text{話題性} \\ \text{閲覧数} < \text{第三四分位数} + \text{四分位範囲} \times 1.5 & \text{定常性} \end{cases}$$

各俳優について，日ごとの閲覧数のデータを取得したのち，俳優ごとに第三四分位数と四分位範囲を求める．その後，第三四分位数に四分位範囲を 1.5 倍したものを足し合わせ，その値以上の場合の閲覧数を話題性によるもの，未満の場合の閲覧数を定常性によるものとする．したがって，俳優ごとに話題性と定常性の基準は異なっている．この処理を行うことによって，閲覧数を話題性によるものと定常性によるものとに分離した．

3.3　定常性・話題性の分布と知名度の関係

各俳優について，閲覧数を話題性による閲覧数と定常性による閲覧数に分離した．その後，それぞれについて全期間に当たる総和を取ったのち，その和を話題性と定常性ごとに全俳優に対する値の平均が 0，分散が 1 となるように標準化し，知名度との関係を図 3,4,5,6 のようにプロットした．図 3,4,5,6 は，それぞれ All，Millennials，Gen X，BabyBoomers の知名度デー

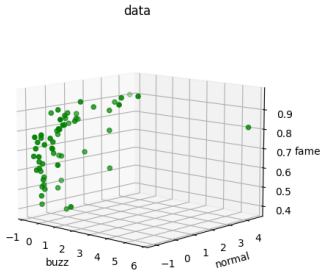

図 3　**話題性・定常性による閲覧数と知名度 (All) の分布．**buzz が話題性による閲覧数，normal が定常性による閲覧数，Fame が知名度 (All) を表す．話題性による閲覧数と定常性による閲覧数はそれぞれ標準化している．各点が一人の俳優を示している．

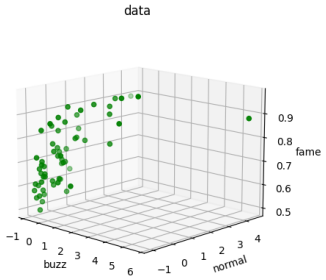

図 4　**話題性・定常性による閲覧数と知名度 (Millennials) の分布．**

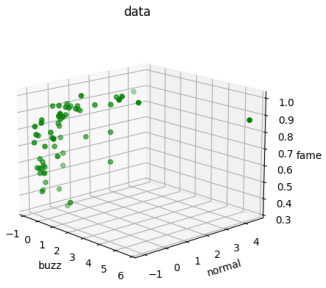

図 5　**話題性・定常性による閲覧数と知名度 (Gen X) の分布．**

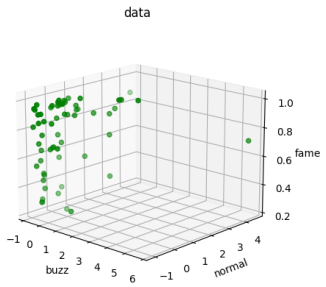

図 6　**話題性・定常性による閲覧数と知名度 (Baby Boomers) の分布．**

タに関するものである．図において，buzz は話題性による閲覧数を，normal が定常性による閲覧数を，Fame が知名度を表す．

図から，すべての世代において，話題性による閲覧数だけで見ると，話題性による閲覧数が同程度の値である狭い範囲内に，知名度が高い俳優から低い俳優までが幅広く分布しているということがわかる．また，Millennials に対する知名度と定常性による閲覧数にはある程度の相関がみられる一方で，Gen X や Baby Boomers など，年齢層の高い世代に対する知名度と定常性による閲覧数にはあまり相関がみられなかった．

4 提案手法

本章では，閲覧数に基づく知名度の推定方法についての提案手法を述べる．知名度の推定には線形回帰を用いる．閲覧数は，定常性による閲覧数と話題性による閲覧数に分割し，一定期間ごとに区切って，期間ごとに平均 0，分散 1 で標準化したものを入力データとした．この時，前述のように，定常性による閲覧数と話題性による閲覧数の情報をどのように組み合わせるかについて，四つの組み合わせ方について比較を行う．さらに，閲覧数の特徴量への変換方法についても三つの方法を比較する．

4.1 閲覧数の期間に基づく分割

知名度は時間の経過とともに累積的に増加していくものであると考えられる．しかし，昔調べたことを多くの人は時間の経過とともに忘れる傾向にある．人物についても，昔調べたことがあったとしても，現在はそのことを忘れてしまっている可能性がある．そこで，昔の閲覧数と最近の閲覧数では知名度に与える影響が異なると考えられるため，一定区間ごとに閲覧数を分割し，閲覧した時期を考慮できるようにした．分割方法は，例えば 10 日間を三分割する際には，[1 ～ 3 日,4 ～ 6 日,7 ～ 10 日] というように，できるだけ均等な間隔となるようにした．

4.2 入力データのパターン

入力データとして，以下の 4 パターンを用いた．

- 定常性による閲覧数
- 話題性による閲覧数
- 定常性による閲覧数 & 話題性による閲覧数
- 合計

「定常性による閲覧数」は，入力データとして定常性による閲覧数のみを用いる．「話題性による閲覧数」は，入力データとして話題性による閲覧数のみを用いる．「定常性による閲覧数 & 話題性による閲覧数」は，定常性による閲覧数と話題性による閲覧数をそれぞれ取得したのち，連結することによって用いる．「合計」は，定常性による閲覧数と話題性による閲覧数を分けずに用いる．

4.3 線形回帰

4.1 節で説明した，分割した各期間をそれぞれ一つの入力次元とし，線形回帰を行う．これを，4.2 節で説明した複数パターンの入力データに対して行う．この時，特徴ベクトルとして，以下の三つを考える．

- $\phi_1(X) = (x_1, x_2, ..., x_h)$
- $\phi_2(X) = (\log(x_1+1), \log(x_2+1), ..., \log(x_h+1))$
- $\phi_3(X) = (x_1^{\frac{1}{1+h}}, x_2^{\frac{1}{(1+h)-1}}, ..., x_h^{1/2})$

ここで，h は期間の数を表す．h が大きくなるほど，その区間は新しいことを意味する．

4.3.1 $\phi_1(X) = (x_1, x_2, ..., x_h)$

線形回帰において重みを学習する際，それぞれの期間における閲覧数を同様にして扱うための特徴ベクトル．重回帰を行っている．

4.3.2 $\phi_2(X) = (\log(x_1+1), \log(x_2+1), ..., \log(x_h+1))$

入力ベクトル X に非線形変換を行うための特徴ベクトル．閲覧数と知名度に非線形な関係性がみられる場合を表現することができる．log の中に 1 加えているのは，x が 0 の時に発散してしまうのを防ぐためである．

4.3.3 $\phi_3(X) = (x_1^{\frac{1}{1+h}}, x_2^{\frac{1}{(1+h)-1}}, ..., x_h^{1/2})$

古い期間になればなるほど，閲覧数の違いが小さくなるようにするための特徴ベクトル．入力データとして「定常性による閲覧数 & 話題性による閲覧数」を使う際には，$\phi_3(X) = (x_1^{\frac{1}{1+h}}, x_2^{\frac{1}{(1+h)-1}}, ..., x_h^{1/2}, x_{h+1}^{\frac{1}{1+h}}, x_{h+2}^{\frac{1}{(1+h)-1}}, ..., x_{2h}^{1/2})$ という形で用いる．ここで，$x_1, \cdots, x_h$ は定常性による閲覧数を，$x_{h+1}, \cdots, x_{2h}$ は話題性による閲覧数を意味する．

5 実　　験

本章では，具体的な実験方法について説明する．

5.1 データセット

3 章で説明したように，知名度の正解データには YouGov-America が提供している，The Most Famous All-time Actors (Q4 2022) から抽出した 61 人の俳優の 4 種類の Fame を用いた．61 人の俳優は，10 位，20 位，30 位，…と 10 位間隔で知名度の順に選択し，そのうちアメリカの俳優のみを取得した．Fame はそれぞれ All, Millennials, Gen X, Baby Boomers となっている．Millenials は 1982-1999 生まれ，Gen X は 1965-1981 生まれ,Baby Boomers は 1946-1964 生まれの世代に対する知名度を表す．また，閲覧回数のデータは，Wikipedia のツールである Pageviews Analysis を用いて取得した．取得する際，オプションで期間は「2015 年 7 月 1 日から 2022 年 9 月 30 日まで」，データ型は「日別」，プロジェクトは「en.wikipedia.org」, プラットフォームは「すべて」，エージェントは「利用者」とした．

5.2 知名度推定の検証方法

知名度推定の精度を検証するために，Leave-One-Out 交差検証を用いた．具体的には，60 人の俳優のデータを用いて線形回帰の学習を行い，残る一人の知名度の推定を行った．この時，推定知名度± 5%の範囲に実際の知名度が入っている場合正しく推定できたとし，入っていなければ失敗したとする．これは，実際に知名度を知りたいと考えた時，知名度の正確な値を知るというよりは，大体の値を知ることができれば十分であると考えられるからである．したがって，知名度の正確な値への近さを表現できるようにするには二乗誤差を用いた方がよいと考え

入力データ	特徴ベクトル	All	Millennials	Gen X	Baby Boomers
定常性による閲覧数	ϕ_1	0.196	0.409	0.114	0.049
	ϕ_2	0.213	0.426	0.196	0.065
	ϕ_3	0.229	0.426	0.196	0.065
話題性による閲覧数	ϕ_1	0.163	0.295	0.147	0.049
	ϕ_2	0.196	0.377	0.131	0.065
	ϕ_3	0.18	0.327	0.098	0.049
定常性 & 話題性による閲覧数	ϕ_1	**0.278**	0.36	0.18	**0.098**
	ϕ_2	0.262	0.426	**0.213**	0.032
	ϕ_3	**0.278**	0.393	0.18	0.081
合計	ϕ_1	0.147	0.426	0.114	0.049
	ϕ_2	0.196	0.409	0.147	0.032
	ϕ_3	0.18	**0.442**	0.147	0.049

表 1　入力データを分割せずに推定を行った実験結果．各要素は正解率を表す．正解率は小数点第四位以下を切り捨て．

られるが，今回の実験ではこのような正解の定義とした．この正解の定義で 61 回の知名度推定を行い，何人の知名度推定を正解することができたのかで評価を行った．

5.3 仮　　説

知名度推定を実施すると同時に，以下の仮説の検証を行う．

（1） 時系列情報を加味することで，知名度推定の精度は向上するか？

（2） 知名度と閲覧数には線形・非線形どちらかの関係があるのではないか？

（3） 定常性と話題性に分割することで，知名度の推定をよりうまく行えるのではないか？

（4） 閲覧数と世代ごとの知名度は関係があるのではないか?

5.4 実験結果と仮説検証

実験は，python3.8.5，scikit-learn1.0.2 の環境で，scikit-learn パッケージに含まれる LinearRegression モデルを用いて行った．閲覧数の期間に基づく分割は最大 15 分割まで行った．実験結果は図 7 から図 22 となっている．図において，赤色の線が ϕ_1，青色の線が ϕ_2, 緑色の線が ϕ_3 を表している．縦軸が線形回帰による知名度推定の精度を，横軸が分割数を表している．閲覧数を期間で分割せずに行った実験結果を表 1 にまとめた．表中の正解率は小数点第四位以下を切り捨てしている．知名度推定の精度は，現状では最大でも五割を超えることができなかったため，閲覧数の情報をそのまま利用する以外の工夫をする必要があることが判明した．以下では，先に述べた仮説を検証していく．

5.4.1 時系列情報を加味することで，知名度推定の精度は向上するか？

図 7 から図 22 にあるように，分割数が増加することによって，知名度推定の精度が増加するといった傾向はあまり見られない．図 8,16,20 のように，入力データや特徴ベクトルによっては分割数を増やすと精度が向上したが，現時点では単純に想定したような効果が出ているとは断言できず，より分析が必要である．したがって，時系列情報を加味することによって，知名度推定の精度が向上するとは言えないということが判明した．

5.4.2 知名度と閲覧数には線形・非線形どちらかの関係があるのではないか？

図 7 から図 22 より，多くの場合で閲覧数の特徴量への変換の仕方の違いによる推定精度の違いは見られなかった．図 8 や図 16，図 20 において，分割数が増えるほど非線形な特徴ベクトルのほうが推定精度が高いという結果が得られたが，現時点では単純に想定したような違いが出ているとは断言できず，より分析が必要である．したがって，知名度と閲覧数には線形・非線形のどちらか一方の関係があるとは言えないということが判明した．

5.4.3 定常性と話題性に分割することで，知名度の推定をよりうまく行えるのではないか？

閲覧時期の分割による知名度推定精度の変化は，現時点ではその効果の有無が断言できないため，分割を行わずに推定を行った結果で分析を行う．表 1 より，入力データとして「定常性による閲覧数 & 話題性による閲覧数」を用いた場合，知名度 All に対してほかの入力データを用いた場合よりも推定精度が高いということが判明した．各世代ごとに比較すると,「定常性による閲覧数 & 話題性による閲覧数」は，知名度 Millennials において「定常性による閲覧数」・「合計」に対して推定精度が同等もしくは劣っている傾向がみられた．一方で，知名度 Gen X と知名度 Baby Boomers においては,「定常性による閲覧数 & 話題性による閲覧数」がそのほかの入力データよりも精度が高い傾向にあるということが判明した．このことより，定常性と話題性に分割し，同時に利用することによって全世代に対する知名度推定の精度は向上するということが判明した．一方で，若い世代の知名度推定となると，分割しない場合及び定常性のみを用いた場合と精度が同等もしくは劣るということが判明した．

5.4.4 閲覧数と世代ごとの知名度は関係があるのではないか？

表 1 より，知名度推定の精度には Millennials > Gen X > Baby Boomers という明確な違いがあるということが判明した．このことより，閲覧数と世代ごとの知名度には関係があるということが分かる．

この理由としては，以下のことが考えられる．対象となる俳優が有名であればあるほど，若い人はその俳優について検索する可能性が高くなり，したがって検索数が増え，Wikipedia の閲覧数も増加する．その結果，知名度と閲覧数の相関が高くなり，知名度推定の精度が高くなると考えられる．一方で，高齢になればなるほど，対象となる俳優を知っていたとしてもその俳優について検索しなくなる．その結果，検索数が少なくても知名度が高いということが発生し，知名度と閲覧数の相関が小さくなり，知名度推定の精度が低くなる．

6 ま と め

本研究では，Wikipedia ページの閲覧回数に基づく知名度推定の手法を提案した．ユーザの検索行動には定常性と話題性が存在すると仮定し，定常性と話題性に分離した閲覧数をもとに知名度の推定を行った．定常性による閲覧数と話題性による閲覧数の複数パターンの組み合わせを入力データとし，正解データとして 4 種類の知名度を用いて線形回帰を行った．この結果,

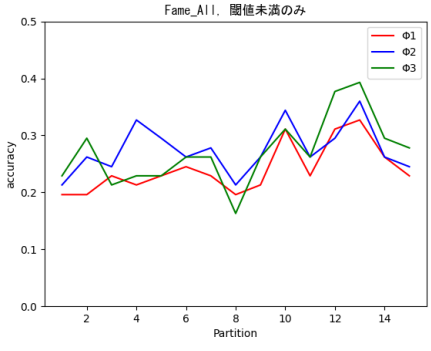

図 7　**知名度:All, 定常性による閲覧数.** 横軸が分割数，縦軸が精度を表す．赤が ϕ_1，青が ϕ_2，緑が ϕ_3 による推定結果を表す．

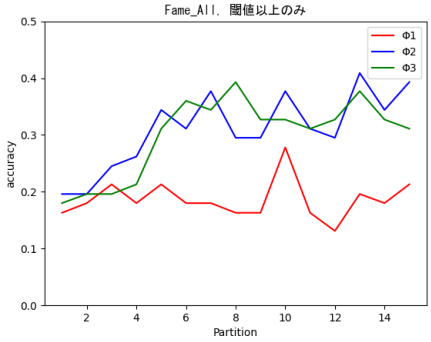

図 8　**知名度:All, 話題性による閲覧数**

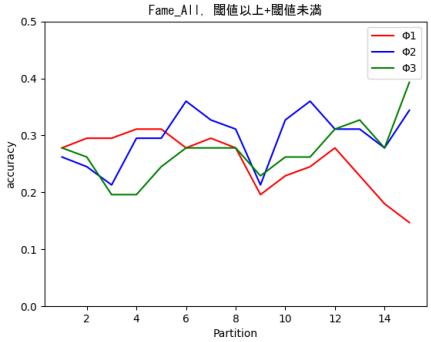

図 9　**知名度:All, 定常性による閲覧数 & 話題性による閲覧数**

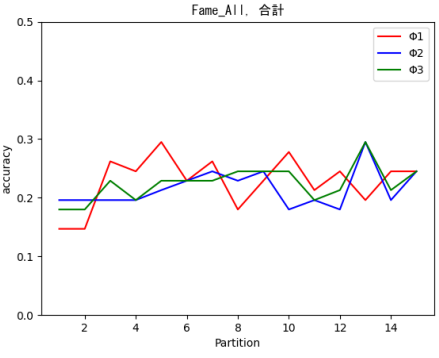

図 10　**知名度:All, 合計**

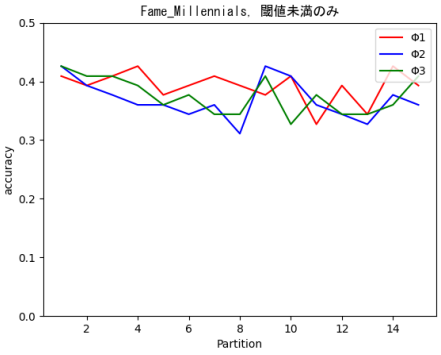

図 11　**知名度:Millennials, 定常性による閲覧数**

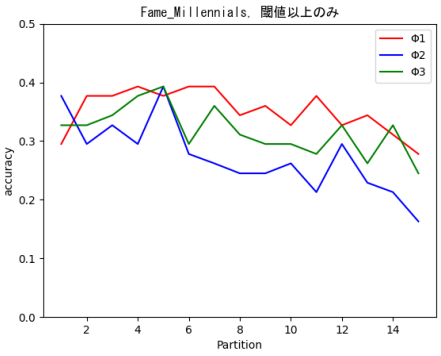

図 12　**知名度:Millennials, 話題性による閲覧数**

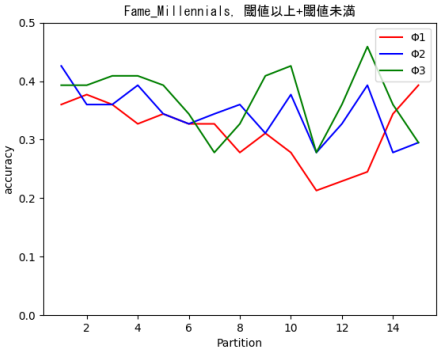

図 13　**知名度:Millennials, 定常性による閲覧数 & 話題性による閲覧数**

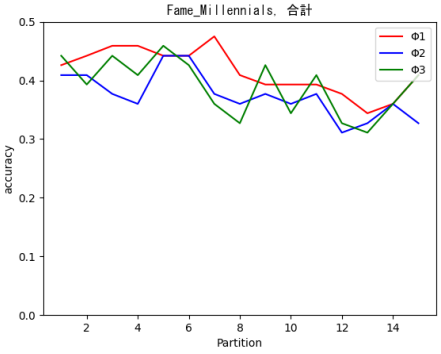

図 14　**知名度:Millennials, 合計**

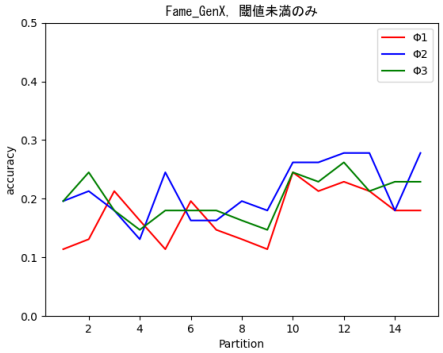

図 15　知名度:Gen X, 定常性による閲覧数

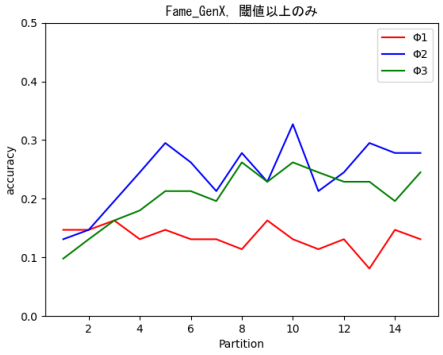

図 16　知名度:Gen X, 話題性による閲覧数

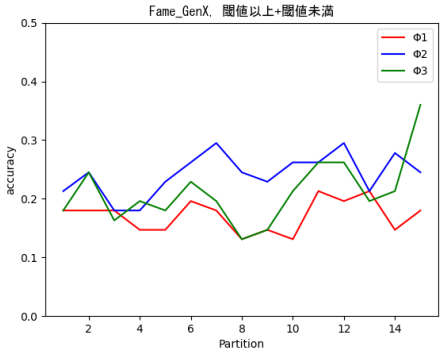

図 17　知名度:Gen X, 定常性による閲覧数 & 話題性による閲覧数

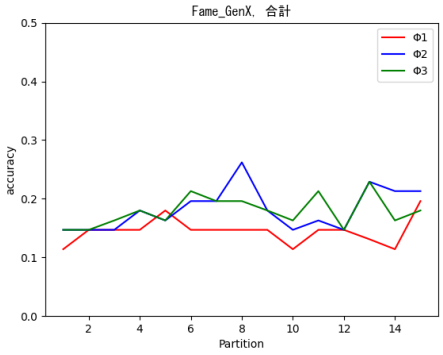

図 18　知名度:Gen X, 合計

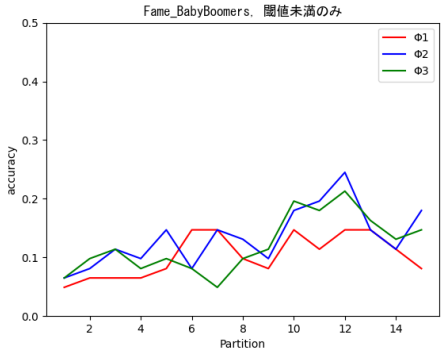

図 19　知名度:Baby Boomers, 定常性による閲覧数

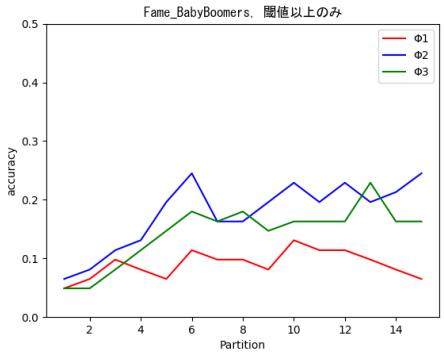

図 20　知名度:Baby Boomers, 話題性による閲覧数

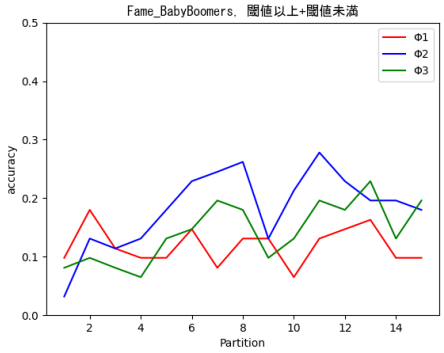

図 21　知名度:Baby Boomers, 定常性による閲覧数 & 話題性による閲覧数

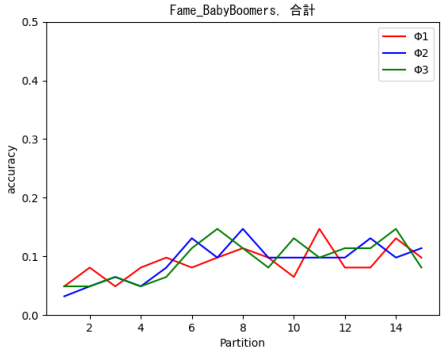

図 22　知名度:Baby Boomers, 合計

閲覧時期の時系列情報を加味することによって知名度推定の精度が向上するとは言えないということ，知名度と閲覧数には線形・非線形どちらか一方の関係があるとは言えないということが明らかになった．一方で，定常性と話題性に分離し同時に用いることによって全世代に対する知名度推定の精度が向上すること，若い世代ほど閲覧数と知名度に高い関係があるということが明らかになった．とはいえ，知名度推定の精度は最大でも 4 割強ほどしかなかったため，閲覧数の情報のみを用いて単純な手法で知名度の推定を行うことには限界があるということが明らかになった．

今後の研究としては，閲覧数以外の情報，例えば有名人のジャンルや年齢などの属性を考慮に入れるなどして実験を行うこと，話題性と定常性に分割することで精度が向上する原因を調査することなどを計画している．また，データ数も増やして実験を行いたいと考えている．

謝　　辞

本研究は JSPS 科研費 21H03446 の助成を受けたものです．

文　　献

[1] 小紫弘貴, 田島敬史. カテゴリごとの偏りを考慮した検索結果からの知名度推定. 第 7 回データ工学と情報マネジメントに関するフォーラム (DEIM 2015) 論文集. 電子情報通信学会, March 2015.

[2] 吉田光男, 荒瀬由紀, 角田孝昭, 山本幹雄. 検索頻度推定のための wikipedia ページビューデータの分析. 人工知能学会全国大会論文集, Vol. JSAI2015, pp. 2I11–2I11, 2015.

[3] Masahiro Inoue and Keishi Tajima. Temporal analysis of supply and demand of topics on the web. In *Proc. of ACM Web Science Conference*, pp. 143–144, June 2019.

[4] Qiang Ma and Masatoshi Yoshikawa. Ranking people based on metadata analysis of search results. In Sven Hartmann, Xiaofang Zhou, and Markus Kirchberg, editors, *Web Information Systems Engineering – WISE 2008 Workshops*, pp. 48–60, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[5] Masaru Kitsuregawa, Takayuki Tamura, Masashi Toyoda, and Nobuhiro Kaji. Socio-sense: A system for analysing the societal behavior from long term web archive. In *Progress in WWW Research and Development: 10th Asia-Pacific Web Conference, APWeb 2008, Shenyang, China, April 26-28, 2008. Proceedings 10*, pp. 1–8. Springer, 2008.

[6] 喜連川優, 豊田正史, 田村孝之, 鍛治伸裕, 今村誠, 高山泰博, 藤原聡子ほか. 学と産の連携による基盤ソフトウェアの先進的開発: 10. socio sense: 過去 9 年に及ぶ web アーカイブから社会の動きを読む. 情報処理, Vol. 49, No. 11, pp. 1290–1296, 2008.

[7] Suppawong Tuarob, Conrad S. Tucker, Marcel Salathe, and Nilam Ram. Modeling individual-level infection dynamics using social network information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, p. 1501–1510, New York, NY, USA, 2015. Association for Computing Machinery.

[8] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, p. 851–860, New York, NY, USA, 2010. Association for Computing Machinery.

[9] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. Rapid assessment of disaster damage using social media activity. *Science Advances*, Vol. 2, No. 3, p. e1500779, 2016.

[10] YouGovAmerica. The most popular all-time actors. `https://today.yougov.com/ratings/entertainment/popularity/all-time-actors/all.html`.