Wikipedia閲覧数を用いた書籍売上予測を通じた Wikipedia閲覧数への検索エンジンによる影響の分析

櫻井　慎也†　　田島　敬史††

† 京都大学工学部情報学科　〒 606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †sakurai@dl.soc.i.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　本研究では Wikipedia のページビュー数を検索エンジンの検索順位, スニペットの長さで補正して書籍の売上を予測する手法を開発することを通じて,Wikipedia のページビュー数に対する検索エンジンの影響の有無について分析する. 提案手法では,Wikipedia のページビュー数を用いる際に, そのページの検索エンジンにおける順位やスニペットの長さを考慮することで予測精度を改善する. Wikipedia のページビュー数に検索エンジンの検索順位やスニペットの長さを加味することによって書籍の売上の予測精度を改善することができたことから,Wikipedia ページビュー数に対して検索エンジンの検索順位やスニペットの長さが影響を与えているということが分かった.

キーワード　Wikipedia, 回帰分析, Web 情報, 検索順位, スニペット

1. はじめに

書籍市場のマーケティング活動において各書籍の売上を予測することは非常に重要なことである. 各書籍が現在どの程度売れていて, そして今後どれくらい売れるのかを予測することで企業の利益を最大化することが可能になる.

書籍の売上を予測する手段としては Wikipedia の該当作品のページビュー数を用いる方法がある. しかし Wikipedia のページビュー数には検索エンジンによるバイアスが存在すると考えられる. 例えば,Wikipedia のページの検索順位が低い場合, 検索者は Wikipedia のページよりも検索順位が上位のページにアクセスしやすいと考えられ, 検索順位が低いほどページビュー数が少なくなると考えられる.

本研究ではこの Wikipedia のページビュー数に加えて, 該当ページの Google 検索による検索順位やスニペットの長さを考慮することで書籍売上の予測精度の向上を図り, これを通じて Wikipedia のページビュー数に対して検索エンジンがどのような影響を与えているのかを考察する. 本研究の概念図を図 1 に示す.

本研究の新規性と有用性は, 検索エンジンの検索順位やスニペットの長さというデータを用いて書籍の売上を予測するという点にある. これまでの研究として, 時系列データに対するアプローチから売上を予測するという研究はあったが, 本研究では時系列データに対するアプローチは用いない. また,Wikipedia のページビュー数は書籍の売上だけではなく, 映画の興行収入や CD の売上など, 様々な対象に対しても用いることができると考えられる.

本論文は以下のように構成される. まず 2 章では, 本研究のテーマと関連性の高いテーマについて扱っている研究を紹介する. 3 章では, 本研究で用いる Web 上のデータの収集方法について述べる. 4 章では, 収集したデータを用いて回帰分析を行い, 本研究の提案手法について述べる. 5 章では, 本研究で提案している手法が有用であるのかを評価するための実験について述べる. 6 章では, 評価実験の結果とそれに対する考察を述べる. 7 章では, 本研究の結論を述べる.

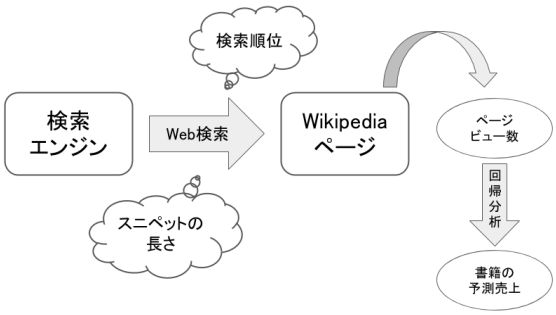

図 1　本研究の概念図

2. 関連研究

本章では, 本研究のテーマと関連性の高いテーマについて扱っている研究を紹介する.

2.1 Web 上の情報を用いた予測モデル

Web 上の情報を用いて売上を予測する研究は多く存在する. 例えば, 保住ら [1] は検索エンジンにおける検索回数や Twitter 上のツイート回数,Wikipedia の編集回数などの素性を用いてコンテンツの消費トレンドを予測するシステムを構築している. また, 他にも Asur ら [2] による Twitter 上のツイート数から映画の売上を予測するという研究や野中ら [3] による Web 上のレビュー情報から自動車の売上を予測するという研究なども行われている. しかし, これらの研究では検索エンジンにおける検索順位という素性は用いられておらず, 筆者の知る限りでは検

索エンジンの検索順位を用いて売上を予測するという研究は他に存在しない.

2.2 時系列データに対するアプローチを用いた売上予測

売上データの推移は時系列データとして捉える事ができるため, 時系列データに対するアプローチを用いて売上を予測する研究も多く存在する. 例えば, 小柳ら [4] はロジスティック方程式を用いて CD の売上予測式を立て,CD の最終売上を予測している. また, 橋本ら [5] による季節による売上の変動を加味して大型小売店における売上を予測するという研究も存在する. しかし, 時系列データに対するアプローチを用いる場合, 映像化等によって急激に売上が変動した場合については対応できないため, 本研究のように時系列データに依存しない手法も有用であると言える.

2.3 検索順位のバイアスに関する研究

検索エンジンの検索順位によるバイアスに関連する研究を紹介する. Chapelle ら [6] は URL が検索順位の下位に現れるサイトは仮に関連性が高かったとしてもユーザーにクリックされにくくなるというバイアスが存在するということを指摘し, ユーザーのクリック履歴を用いた動的ベイジアンネットワークを利用することによってこのような検索順位の低さに起因するバイアスを軽減することができるということを示している.

Joachims ら [7] はアイトラッキングを用いてユーザーの意思決定の過程を観察し,Web 検索エンジンのクリック率は有用な情報であるが検索順位によるバイアスが含まれているということを明らかにした. 本研究も検索エンジンの検索順位によるバイアスについて着目しているが, 書籍の売上という実際のデータとの関連性を言及しているという点で他の研究とは異なっている.

2.4 検索順位以外のバイアスに関する研究

検索エンジンにおける検索順位以外のバイアスについて言及している研究も多く存在する. Ieong ら [8] は検索結果のページの属するドメインがユーザーの選択に影響を与えているということを実験を通じて明らかにしている. Yue ら [9] は検索結果のタイトルやスニペットの内容によってユーザーの選択にバイアスがかかっているということを明らかにしている.

2.5 Wikipedia のページビューデータに関する研究

Wikipedia のページビューデータを用いた研究も多く存在する. 吉田ら [10] は検索頻度を推定するために Wikipedia のページビュー数を利用することを考え,Wikipedia のページビュー数と検索頻度との類似性について調査している. Mestyan ら [11] は Wikipedia のページビュー数を用いて映画の流行を予測することを試みている.

3. Web 上のデータの収集

本章では, 検索エンジンの検索順位が Wikipedia のページビュー数と原作書籍の売上げランキングに影響を与えていることを検証するために収集した Web 上のデータの概要を述べる.

本研究では, 書籍の推定売上のデータを書籍ランキングデータベース (注1)の POS 週間文庫ランキングから取得し, 検索エンジンの検索順位とスニペットのデータを Google 検索を用いて取得し,Wikipedia のページビュー数を MediaWiki の PageviewAPI を用いて取得した.

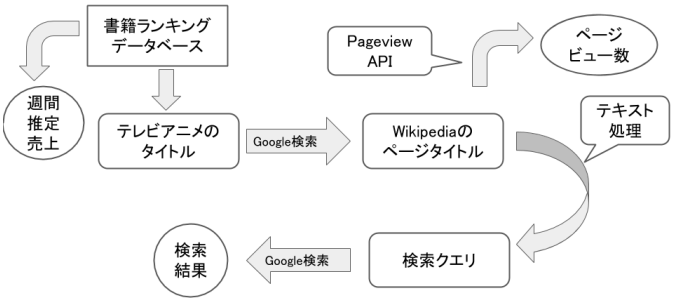

図 2 データ収集の大まかな流れ

3.1 収集対象

本研究の収集対象となった書籍は以下の条件を全て満たすものである.

- 2015 年 9 月 1 日から 2017 年 12 月 31 日までの間に公開されたテレビアニメの原作書籍である
- 同一シリーズの書籍の中で最も発売日が古い
- アニメ作品の放送期間中に POS 週間文庫ランキングの 500 位以内にランクインしている

3.2 収集期間

集計期間は上記の条件にあるように 2015 年 9 月 1 日から 2017 年 12 月 31 日までの 2 年間であり, この期間に放送された各テレビアニメ作品の放送期間それぞれ 12 週間に渡って取得した.

3.3 書籍の推定売上

書籍の推定売上は書籍ランキングデータベースの POS 週間文庫ランキングから取得した. このとき, ランキング上位 500 位以内に入っていなければランキング圏外となり推定売上を算出することができないため, 売上ランキング上位 500 位以内に入っていない週については対象外とした.

3.4 検索クエリ

検索クエリは検索順位やスニペットを決定づける大きな要因の一つである. 本研究では適切な検索順位やスニペットを得るために, 検索クエリを適切な形になるように調整した. 検索クエリの調整を含めたデータ収集の大まかな流れは図 2 のようになる.

まず, 書籍ランキングデータベースから取得した,3.1 節で示した条件に合致するテレビアニメのタイトルを検索クエリとして Google 検索を行なった. この検索によって該当作品に関連する Wikipedia のページが得られるが, テレビアニメのタイトルをそのまま検索クエリとして利用するとタイトルに余分なものが含まれることがあった.

例えば, アニメのタイトル “ キノの旅 the Beautiful World

(注1) : http://book-rank.net/index.cgi

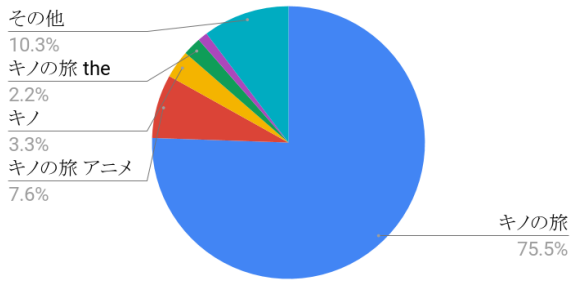

図 3　“キノの旅”のアニメ公式サイトへの検索クエリごとの流入割合

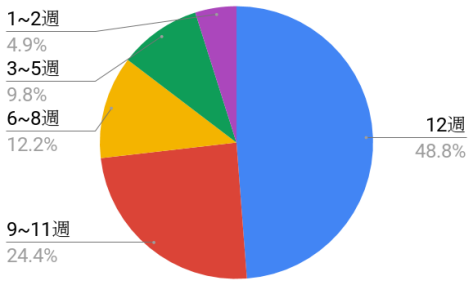

図 4　各書籍の取得できた週間推定売上の数

the Animated Series”である作品について考えると, 検索クエリとして使用するのはアニメのタイトルそのままよりも“キノの旅”という検索クエリを用いて検索されることの方が一般的であると考えられる. 実際に,“キノの旅”のアニメ公式サイト(注2)への検索ワードごとの割合を見ると図 3 のようになる. 検索ワードごとの割合は SimilarWeb(注3)を用いて取得した.

このように適切な検索クエリを設定するために, 本研究では Wikipedia のタイトルを利用した. 具体的には, 先述のようにテレビアニメのタイトルを検索クエリとして検索した後, この検索によって得られた該当コンテンツに関する Wikipedia のページのタイトルを取得し, この Wikipedia のページのタイトルに対して前処理を行なった. このときの前処理とは以下の 3 つの処理を指す.

（1） タイトルの末尾にある“- Wikipedia”または“- ウィキペディア”を除去する

（2） タイトルの末尾にある“ (漫画)”や“ (小説)”など, ページの曖昧さ回避のための文字列を除去する

（3） タイトル中にある“-”に囲まれた文字列を除去する

3 つ目の処理は, 例えば“BORUTO -ボルト- -NARUTO NEXT GENERATIONS-”のように長いサブタイトルが含まれている場合にサブタイトルを除去するために行っている. これらの前処理を行なって最終的に得られた文字列を検索クエリとして設定した.

3.5　検索順位とスニペット

前節の手順によって検索クエリの調整を行なった後,Google 検索を用いて検索エンジンにおける Wikipedia のページの検索順位とスニペットの文字数を取得した. 検索順位は最上位に表示されるものから順に 0,1,2,... となるようにした. 検索順位とスニペットは 2018 年 1 月 26 日時点のものを取得している.

3.6　Wikipedia のページビュー数

上述の検索によって得られた Wikipedia のページの一日ごとのページビュー数を MediaWiki の PageviewAPI を用いて取得し,POS 週間文庫ランキングの集計期間に合わせて週間のページビュー数を算出した.

3.7　デ ー タ 数

以上の方法によって得られた書籍の数は 41 冊, 週間推定売上と Wikipedia のページビュー数のペアは 378 組となった. なお, 週間推定売上とページビュー数のペアが 41(冊) × 12(週間)=492(組) よりも小さいのは週間売上ランキングが圏外で, 週間推定売上を算出できない場合を除いているためである. 各書籍の取得できた週間推定売上の数は図 4 のようになった. 図 4 から 12 週間分全ての売上を取得できたものから 1,2 週間分しか取得できなかったものまであることが分かる.

(注2) : http://www.kinonotabi-anime.com/
(注3) : https://www.similarweb.com/ja

4.　提 案 手 法

本章では検索エンジンから得られる情報を用いて Wikipedia のページビュー数を補正することで書籍の売上予測の精度を向上させることを考える. 本章における回帰分析では 3 章で収集した 378 組のデータ全てを用いて書籍ごとの leave-one-out 交差検証を行なった. leave-one-out 交差検証については 5.1.2 で改めて詳しく述べる.

以下の (1) 式から (9) 式において書籍 i の j 週目の売上の予測値を $s\hat{i}_{,j}$,Wikipedia のページビュー数を $p_{i,j}$ とし, 回帰式の回帰係数を a, 切片を b とする.　また c は $0 < c < 4.0$ の範囲で回帰分析の二乗平均平方根誤差 (RMSE 値) が最小となるものである.　RMSE 値の計算方法については 5.1.3 で改めて詳しく述べる.

4.1　Wikipedia の検索順位で補正する手法

本節では,Wikipedia の検索順位を用いてページビュー数を補正することで書籍の売上予測精度を向上させることを考える. 検索順位を求めるための検索エンジンには Google 検索を利用した. また, 以下の式において書籍 i に関する Wikipedia ページの検索順位を r_i とする.

4.1.1　仮　　説

まず本研究を行うにあたって立てた仮説について説明する. 本研究では Wikipedia のページ閲覧数を用いて書籍の売上を予測する際に, 検索エンジンにおける検索順位を利用することで Wikipedia のページ閲覧数に対する検索順位の影響を調べることを目的としている.　本研究では検索順位が低い, つまり Wikipedia のページが検索結果のより下位に現れるほど, 実際の書籍の売上に対して Wikipedia のページ閲覧数が少なくなるという仮説を立てた.　この仮説は, 一般的に検索エンジンの利用者は検索結果を上位から順番に確認していくため,Wikipedia のページの検索順位が低いと検索エンジンで検索をしたが Wikipedia のページよりも上位のページを見ただけで Wikipedia のページを見ないという潜在的な検索数をカウント

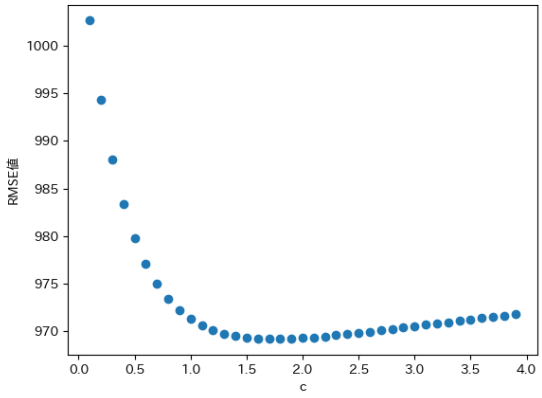

図 5　検索順位の対数関数を用いた回帰式の RMSE 値

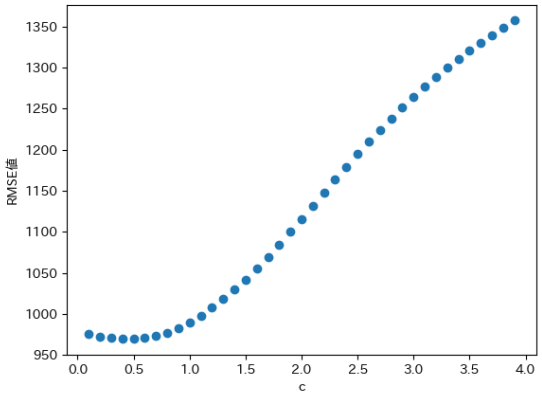

図 6　検索順位を底とする指数関数を用いた回帰式の RMSE 値

することができないという根拠に基づいている．この仮説から，本研究では検索順位 r_i が大きいほど潜在的な検索数を含めたページビュー数が大きくなるように回帰式を決定した．

4.1.2　検索順位の対数関数を用いた回帰式

まず (1) 式のように検索順位の対数関数で Wikipedia のページ閲覧数を補正した回帰式を用いることで書籍の売上を予測する場合を考える．(1) 式の対数の真数が $(1+c+r_i)$ となっているのは $r_i=0$ のとき $log_{1+c}(1+c+r_i)=1$ とするためである．

$$\hat{s_{i,j}} = ap_{i,j} log_{1+c}(1+c+r_i) + b \tag{1}$$

例として，(1) 式の c を変化させながら回帰分析を行い，RMSE 値をプロットすることで RMSE 値が最小となる c を求める．c を $0 < c < 4.0$ の範囲で変化させながら RMSE 値をプロットすると図 5 のようになる．RMSE 値が最小となる c は $c = 1.8$ であり，そのときの RMSE 値は 969 であった．

4.1.3　検索順位を底とする指数関数を用いた回帰式

次に (2) 式のように検索順位を底とする指数関数で Wikipedia のページ閲覧数を補正した回帰式を用いることで書籍の売上を予測する場合を考える．

$$\hat{s_{i,j}} = ap_{i,j}(r_i+1)^c + b \tag{2}$$

c を $0 < c < 4.0$ の範囲で変化させながら RMSE 値をプロットすると図 6 のようになる．RMSE 値が最小となる c は $c = 0.5$ であり，そのときの RMSE 値は 969 であった．

4.1.4　検索順位を指数とする指数関数を用いた回帰式

次に (3) 式のように検索順位を指数とする指数関数で Wikipedia のページ閲覧数を補正した回帰式を用いることで書籍の売上を予測する場合を考える．

$$\hat{s_{i,j}} = ap_{i,j}(1+c)^{r_i} + b \tag{3}$$

c を $0 < c < 4.0$ の範囲で変化させながら RMSE 値をプロットすると図 7 のようになる．RMSE 値が最小となる c は $c = 0.2$ であり，そのときの RMSE 値は 967 であった．

4.1.5　検索順位別のクリック率を利用した回帰式

次に，Web 上に公開されている検索順位別のクリック率のデータを利用して Wikipedia のページ閲覧数を補正した回帰式を用いることで書籍の売上を予測する場合を考える．検索順位別のクリック率のデータは 2014 年に NetBooster が公開した調査結果 (注4) を利用した．30 位までの検索順位別のクリック率は図 8 のようになっている．検索順位が r 位のときのクリック率を $f(r)$ として，(4) 式のような回帰式を用いて回帰分析を行う．

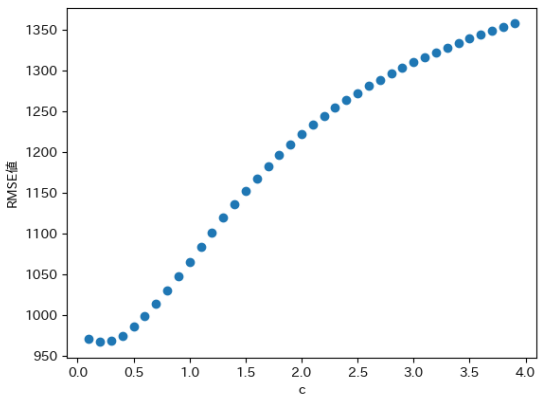

図 7　検索順位を指数とする指数関数を用いた回帰式の RMSE 値

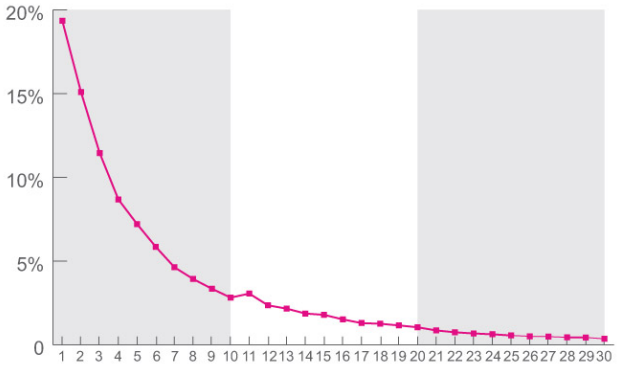

図 8　上位 30 位までの検索順位別クリック率

$$\hat{s_{i,j}} = \frac{ap_{i,j}}{f(r_i)} + b \tag{4}$$

(注4)：http://www.netbooster.co.uk/one-click-curve-to-rule-them-all/

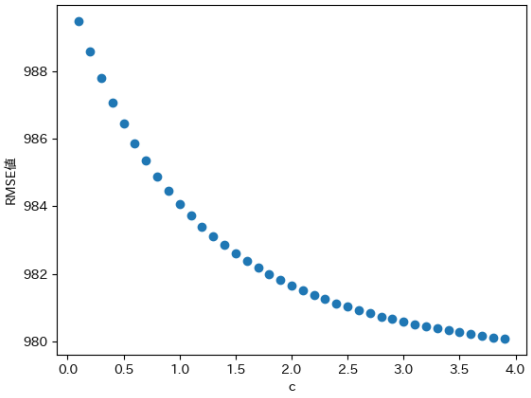

図 9　スニペットの文字数の対数関数を用いた回帰式の RMSE 値

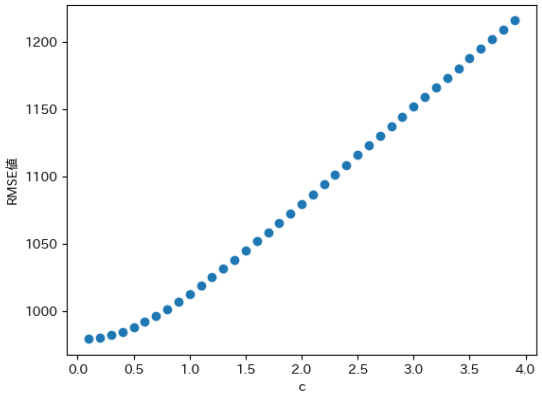

図 10　スニペットの文字数を底とする指数関数を用いた回帰式の RMSE 値

4.2　スニペットの文字数で補正する手法

検索エンジンによる検索結果には, そのページのうち検索クエリに関連する部分の一部を抽出したものが含まれる.　これをスニペットという. 本章ではスニペットの文字数を用いて Wikipedia のページビュー数を補正することで書籍の売上予測の精度を向上させることを考える. 以下の式について,Google 検索における書籍 i に関する Wikipedia ページのスニペットの文字数を l_i とする.

4.2.1　仮　　説

まずスニペットの文字数でページビュー数を補正するに当たって立てた仮説について説明する. スニペットは前述のように検索クエリに関連した部分の一部を抽出したものであり, 一般的にスニペットが長いほど検索クエリに関連した部分が多く含まれていると考えられる. また Yue ら [9] の研究においてもスニペットが長いほどユーザーにとって魅力的なページであり, クリック率も高くなるということが示されている.

しかし一方で, スニペットが長くなればなるほどスニペットに含まれる情報も多くなるため, 検索ユーザーは求めていた情報をスニペットの内容から得られる確率も高くなる. このことからスニペットが長くなるほど検索エンジンから Wikipedia のページへのページ遷移を行わないユーザーの数も多くなると考えられる. つまり, スニペットが長いほど Wikipedia のページビュー数に表れていない潜在的な検索数が多くなるという仮説が立てられる. この仮説から, 本研究ではスニペットの文字数 l_i が大きいほど潜在的な検索数を含めたページビュー数が大きくなるように回帰式を決定した.

4.2.2　スニペットの文字数の対数関数を用いた回帰式

(6) 式のようにスニペットの文字数の対数関数で Wikipedia のページビュー数を補正した回帰式を用いることで書籍の売上を予測する場合を考える.

$$\hat{s_{i,j}} = ap_{i,j} log_{1+c}(1 + c + \frac{l_i}{100}) + b \tag{5}$$

c を $0 < c < 4.0$ の範囲で変化させながら RMSE 値をプロットすると図 9 のようになる. RMSE 値が最小となる c は $c = 3.9$ であり, そのときの RMSE 値は 980 であった.

4.2.3　スニペットの文字数を底とする指数関数を用いた回帰式

(6) 式のようにスニペットの文字数を底とする指数関数で Wikipedia のページビュー数を補正した回帰式を用いることで書籍の売上を予測する場合を考える.

$$\hat{s_{i,j}} = ap_{i,j}(\frac{l_i}{100})^c + b \tag{6}$$

c を $0 < c < 4.0$ の範囲で変化させながら RMSE 値をプロットすると図 10 のようになる. RMSE 値が最小となる c は $c = 0.1$ であり, そのときの RMSE 値は 979 であった.

4.2.4　スニペットの文字数を指数とする指数関数を用いた回帰式

(7) 式のようにスニペットの文字数を指数とする指数関数で Wikipedia のページビュー数を補正した回帰式を用いることで書籍の売上を予測する場合を考える.

$$\hat{s_{i,j}} = ap_{i,j}(1 + c)^{\frac{l_i}{100}} + b \tag{7}$$

c を $10 < c < 40$ の範囲で c を変化させながら RMSE 値をプロットすると図 11 のようになる. RMSE 値が最小となる c は $c = 0.1$ であり, そのときの RMSE 値は 981 であった.

5.　評 価 実 験

本章では 4 章で示した回帰式について評価実験を行い, 本研究の提案手法の有用性を検証する.

5.1　実 験 方 法

5.1.1　データセットの分割

本研究では評価結果の有意性を確かめるために, データセットの分割を行なった. 分割方法を図 12 に示す.

本研究で収集したデータセットは 3.1 節に示した条件を満たす 41 冊の書籍のそれぞれの 12 週間にわたる売上とページビュー数のペア 378 組であるが, この 378 組のデータセットを 1 週間ごとに 12 個のデータ集合に分割し, それぞれのデータ集合に対して後述の leave-one-out 交差検証による評価を行なった.

このようなデータセットの分割を行なった理由は, 本研究で収集した推定売上のデータには 41 冊の書籍ごとに 12 週間分

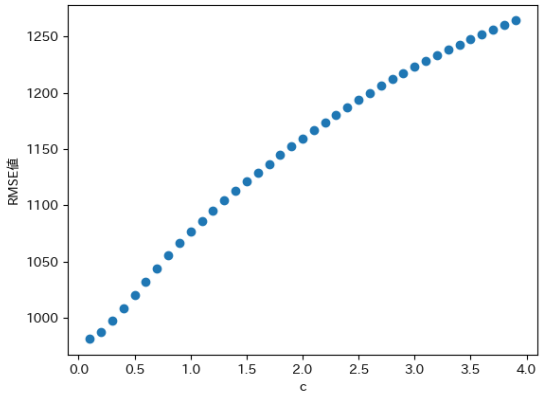

図 11　スニペットの文字数を指数とする指数関数を用いた回帰式の RMSE 値

図 12　データセットの分割方法

を取得できたものから 1 週間分しか取得できなかったものまであり, これらを分割せずに用いた場合, 回帰分析に用いるデータセットの中で各書籍の占める割合が異なってしまうからである. 本研究は検索順位やスニペットの長さというデータを用いており, これらは各書籍と一対一で対応しているが売上やページビュー数とは一対多の対応となっているため, データセットの中で各書籍の占める割合が異なる場合それぞれの検索順位やスニペットの長さの持つ重みに偏りが生じてしまう. 週ごとにデータセットを分割することで各書籍のデータが各データ集合に最大 1 つまでしか現れなくなるため, 検索順位やスニペットの長さの持つ重みの偏りを除去することができる.

5.1.2　交 差 検 証

本研究では書籍単位の leave-one-out 交差検証を行った.leave-one-out 交差検証は収集したデータセットのうち一つを回帰式を評価するためのテストデータとし, 残りのデータセットを全て回帰分析を行うための訓練データとするという分け方を全てのデータセットに対して行ってから回帰式の評価を行うという手法である. 本研究ではこの leave-one-out 交差検証を書籍単位で行い, 書籍 i のデータをテストデータとして書籍 i 以外の全ての書籍のデータを訓練データとするという方法を $i = 1, 2, ..., 41$ について行なった. 書籍単位の leave-one-out 交差検証の流れを図 13 に示す.

5.1.3　二乗平均平方根誤差

本研究では 4 章で示した式を回帰式として 3 章で収集したデータセットを用いて leave-one-out 交差検証を行い, 二乗平均平方根誤差 (RMSE 値) を用いて回帰式の評価を行った. 売上

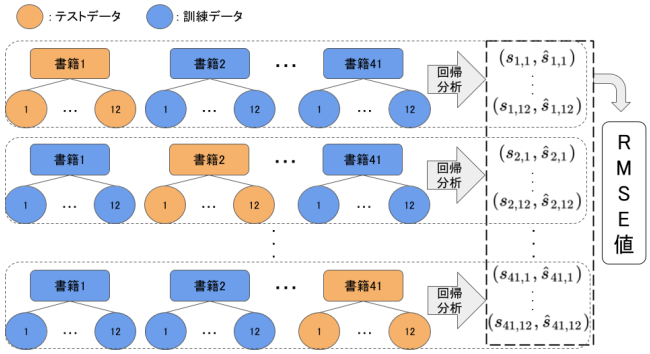

図 13　書籍単位の leave-one-out 交差検証

表 1　検索順位を用いた手法の RMSE 値 (データ分割なし)

手法	RMSE 値
ベースライン	979
4.1.2 の手法	969
4.1.3 の手法	969
4.1.4 の手法	967
4.1.5 の手法	968

の予測値を $\hat{s_i}$, 実際の値を s_i, データ数を n とすると RMSE 値は (8) 式で計算することができる.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(\hat{s_i} - s_i)^2} \tag{8}$$

RMSE 値は回帰式から予測された予測値と実際の値との誤差の二乗の平均値の平方根を取ったものであり,RMSE 値が小さいほど誤差が少なく優れた回帰式であると言える.

5.2　ベースライン

提案手法と比較するためのベースラインについて述べる. 提案手法の有用性を検証するために,Wikipedia のページビュー数を補正せずに線形単回帰分析を行なったものをベースラインとして採用した. ベースラインの回帰分析に用いる回帰式は (9) 式となる.

$$\hat{s_{i,j}} = ap_{i,j} + b \tag{9}$$

6.　実 験 結 果

6.1　データを分割しない場合

まずデータを週ごとに分割せず, 書籍ごとの leave-one-out 交差検証により RMSE 値を求める.

6.1.1　検索順位を用いた手法の評価結果

検索順位を用いて Wikipedia のページビュー数を補正した,4.1 節の各提案手法について RMSE 値を求めると表 1 のようになる.

表 1 を見ると,4.1 節の各提案手法全てにおいてベースラインよりも RMSE 値が小さくなっており, 検索順位を用いた手法が有効に働いていることがわかる. 特に 4.1.4 の手法は最も RMSE 値が小さくなっており,4.1 節の手法の中で最も有用であると考えられる.

表 2　スニペットの長さを用いた手法の RMSE 値 (データ分割なし)

手法	RMSE 値
ベースライン	979
4.2.2 の手法	980
4.2.3 の手法	979
4.2.4 の手法	981

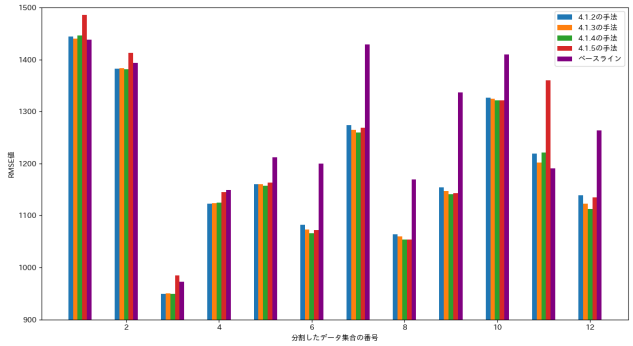

図 14　検索順位を用いた手法の RMSE 値 (データ分割あり)

表 3　検索順位を用いた手法の RMSE 値 (データ分割あり)

データ集合	1	2	3	4	5	6	7	8	9	10	11	12	平均
ベースライン	**1438**	1394	973	1149	1212	1200	1429	1169	1337	1410	**1191**	1264	1263
4.1.2 の手法	1444	1382	**949**	**1123**	1160	1082	1274	1064	1154	1327	1219	1139	1193
4.1.3 の手法	1440	1383	950	1124	1160	1073	1265	1060	1147	1325	1202	1123	1187
4.1.4 の手法	1446	**1381**	**949**	1125	**1157**	**1066**	**1260**	**1054**	**1141**	**1322**	1221	**1113**	**1186**
4.1.5 の手法	1486	1413	985	1145	1163	1072	1269	**1054**	1143	**1322**	1360	1135	1212

6.1.2　スニペットの長さを用いた手法の評価結果

スニペットの長さを用いて Wikipedia のページビュー数を補正した,4.2 節の各提案手法について RMSE 値を求めると表 2 のようになる.

表 2 を見ると,4.2 節の手法はベースラインより RMSE 値が小さくなっておらず, スニペットの長さを用いた手法が有効に働いていないことがわかる.

6.2　データを分割した場合

5.1.1 で示したようにデータセットを週ごとに分割してから学習と評価を行い,leave-one-out 交差検証によって RMSE 値を求める場合を考える.

6.2.1　検索順位を用いた手法の評価結果

分割したデータ集合に対して 4.1 節の手法を適用し,RMSE 値を求めると図 14, 表 3 のようになる.

表 3 を見るとデータ集合 1,2,3,11 以外の全てのデータ集合において,Wikipedia のページビュー数を補正せずに回帰分析を行なったベースラインに比べて検索エンジンの検索順位を用いて Wikipedia のページビュー数を補正して回帰分析を行なった提案手法の方が RMSE 値が小さくなっており, 特に 4.1.4 の手法は RMSE 値の平均が最も小さくデータ集合 1,11 以外の全てのデータ集合においてベースラインよりも RMSE 値が小さくなっていることが分かる.

提案手法では検索順位が低いほど潜在的な検索数が多くなり実際の書籍の売上に比べて Wikipedia のページ閲覧数が少なくなるという仮説をもとに, 検索順位が低い, つまり r_i の値が大きいほど潜在的な検索数を含めたページビュー数が大きくなるように回帰式を設定した.　そしてこの回帰式を用いた場合に RMSE 値が小さくなったということから, 本研究の仮説を検証できた.

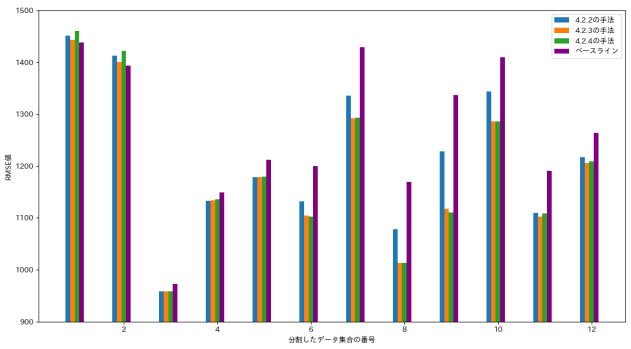

図 15　スニペットの長さを用いた手法の RMSE 値 (データ分割あり)

表 4　スニペットの長さを用いた手法の RMSE 値 (データ分割あり)

データ集合	1	2	3	4	5	6	7	8	9	10	11	12	平均
ベースライン	**1438**	**1394**	973	1149	1212	1200	1429	1169	1337	1410	1191	1264	1263
4.2.2 の手法	1451	1413	**958**	**1133**	**1179**	1132	1336	1078	1228	1344	1110	1217	1215
4.2.3 の手法	1443	1401	**958**	1134	**1179**	1104	**1292**	**1013**	1118	**1286**	**1102**	**1206**	**1186**
4.2.4 の手法	1461	1422	**958**	1136	1180	**1102**	1293	**1013**	**1111**	**1286**	1109	1209	1190

6.2.2　スニペットの長さを用いた手法の評価結果

分割したデータ集合に対して 4.2 節の手法を適用し,RMSE 値を求めると図 15, 表 4 のようになる.

表 4 を見るとデータ集合 1,2 以外の全てのデータ集合において,Wikipedia のページビュー数を補正せずに回帰分析を行なったベースラインに比べて Wikipedia のスニペットの文字数を用いてページビュー数を補正して回帰分析を行なった提案手法の RMSE 値が小さくなっており, 特に 4.2.3 の手法は RMSE 値の平均が最も小さくデータ集合 1,2 以外の全データ集合においてベースラインよりも RMSE 値が小さくなっていることが分かる.

データ分割をしない場合にはスニペットの長さを用いた手法は有効ではなかったが, データ分割をした場合にはスニペットの長さを用いた手法が有効に働いたことから, 各書籍のスニペットの長さの持つ重みを均等化することが有効であったと考えられる.

提案手法ではスニペットの長さが長いほど潜在的な検索数が多くなり実際の書籍の売上に比べて Wikipedia のページ閲覧数が少なくなるという仮説をもとに, スニペットの文字数が多い, つまり l_i の値が大きいほど潜在的な検索数を含めたページビュー数が大きくなるように回帰式を設定した.　そしてこの回帰式を用いた場合に RMSE 値が小さくなったということから, 本研究の仮説を検証できた.

7.　結　　論

本研究では, “ 検索順位が低いほどページビュー数に現れない潜在的な検索数が多くなる ” という仮説と “ スニペットの文字数が長いほどページビュー数に現れない潜在的な検索数が多く

なる ”という仮説を立て, これらの仮説に基づいた回帰式を用いた回帰分析によって書籍の売上予測精度が向上したことから, 仮説の正しさを検証をした.

本研究で利用した検索順位は時間経過による変化を考慮していないが, 一般的に検索順位は時間の経過によって変化すると考えられるため, このような問題を踏まえた上で検索エンジンのページビュー数に対する影響を分析することが今後の課題であると考えられる.

文　　献

[1] 保住純, 飯塚修平, 大澤昇平. Web マイニングを用いたコンテンツ消費トレンド予測システム. 人工知能学会全国大会論文集, Vol. 27, pp. 1–4, 2013.

[2] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pp. 492–499. IEEE Computer Society, 2010.

[3] 野中尚輝, 松尾豊. オンラインレビュー情報の利用による自動車の売上予測手法の提案. 人工知能学会全国大会論文集, Vol. 29, pp. 1–4, 2015.

[4] 小柳文子, 近匡, 綛田侑三. ロジスッティック・モデルから求めた音楽 cd 売上げ予測. 理工学研究報告, 2007.

[5] 橋本郁郎. 大型小売店における売上予測. 1990.

[6] Olivier Chapelle and Ya Zhang. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World Wide Web*, pp. 1–10. ACM, 2009.

[7] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, Vol. 51, pp. 4–11. Acm, 2017.

[8] Samuel Ieong, Nina Mishra, Eldar Sadikov, and Li Zhang. Domain bias in web search. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 413–422. ACM, 2012.

[9] Yisong Yue, Rajan Patel, and Hein Roehrig. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the 19th international conference on World Wide Web*, pp. 1011–1018. ACM, 2010.

[10] 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. 29, pp. 1–4, 2015.

[11] Márton Mestyán, Taha Yasseri, and János Kertész. Early prediction of movie box office success based on wikipedia activity big data. *PloS one*, Vol. 8, No. 8, p. e71226, 2013.