1B2-R-3-5

自己教師あり学習を用いた過去の編集記録からのWikipedia編集回数の予測

Wikipedia Edit Number Prediction from the Past Edit Record Based on Auto-Supervised Learning

吉田　裕[*1]　　大和田　勇人[*2]
Yoshida Yutaka　　Ohwada Hayato

[*1]東京理科大学　大学院　理工学研究科　経営工学専攻
Department of Industrial Administration, Graduate School of Science and Technology, Tokyo University of Science

[*2]同　理工学部　経営工学科
Department of Industrial Administration, Faculty Of Science and Technology

本研究はICDMコンテストに参加し，Wikipedia編集者の編集行動を予測するモデルを構築した．手法には時系列解析におけるARモデルを参考にして，教師あり学習の理論を導入した自己教師あり学習を提案した．特徴空間には文字数や時間などの特徴を投入したが，主に編集回数に関する特徴の取り方を実験で比較した．同時期の関連研究の結果より予測精度が低いが，ICDMコンテストで9位の予測精度であった．多手法との比較により，予測精度を高める方法について考察した．

1. はじめに

本研究は，2011年のICDMで行われたデータマイニングコンテストにおいて提起された問題を研究テーマとしている．

Wikipediaとは非営利組織Wikimedia財団 (WMF) によって運営される開放的かつ共同的な多言語百科事典プロジェクトである．2001年発足以来, インターネットにおける最も大規模で利用者の多い被参照知識となった．

近年,WMFの研究から,Wikipediaの成長が停滞していることがわかっている．2005年以前の英語版Wikipediaでは新規参加者の内最初の編集時点から1年後も活動を続けている者は約40 %であるが, 2007年以降では12-15 %のみである．多くの者がWikipediaコミュニティに参加できないでいることで, プロジェクトの継続が困難になりつつある[*1]．それ故に, 編集者の将来の編集行動を決定する要因を理解することはコミュニティの成長を維持するためにも重要である．2011年のICDMデータマイニングコンテスト[*2]では編集者の将来の編集回数を予測し，高い予測精度のモデルを分析することでこの要因を評価できるとされた．

そこで本研究は, 編集者の将来の編集回数を予測するモデルを過去の編集情報を用いて構築する手法を提案することを目的とする．

2. 問題設定

問題は，2001年1月1日から2010年9月1日までの時系列データである編集記録情報を用いて，各編集者の2010年9月1日から2011年2月1日の5カ月間に行わる編集の回数を予測する回帰モデルを構築することである．データはWMFより英語版Wikipediaにおける記録が提供されたが，予測すべき期間以前のデータをWikipediaから抽出して使用してもよい．

連絡先: 吉田　裕，東京理科大学大学院　理工学研究科　経営工学専攻，千葉県野田市山崎 2641，04(7124)1501，j7412633@ed.tus.ac.jp

∗1 http://strategy.wikimedia.org/wiki/Editor_Trends_Study

∗2 http://www.eecs.wsu.edu/ holder/icdm2011contest/

この問題に対して世界中の研究者が挑戦し，研究成果をWikipediaに提出した．本論文はその研究の一例を示した上で，他の研究との比較を行い，より良い予測方法について考察を行う．比較対象は，高い予測精度を示した [Herring 11] と [Zhang 11] とする．

時系列データから特徴を抽出し予測を行う研究に，[Nakata 10] や，[Crone 10] がある．本研究ではこれらの予測モデルの説明変数の選択方法を参考にした上で，さらに主観的に選択した特徴も投入し，自己教師あり学習のアプローチを提案する．

3. 提案手法

3.1 自己教師あり学習

過去の編集情報から未来の編集情報を予測するモデルを考える．まず編集者単位で考えた時，編集回数を連続値として扱えば，過去の編集回数の線形関数で未来の編集回数を表現する時系列解析における自己回帰 (Auto-Regression) モデルが選択肢として考えられる．これは以下の式で表される．

$$y_q = \beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_{(q-1)} y_{(q-1)} + \epsilon_q \quad (1)$$

ここで，y_t は期間 t における編集回数，$\Phi_0, \Phi_1, \ldots, \Phi_{q-1}$ はパラメータ，ϵ_q は誤差である．このようなモデルを自己回帰モデルと呼び，未来の編集回数 y_q は期間 $1, 2, \ldots, q-1$ における編集回数の線形関数として表現される．すなわちある編集者の未来の編集回数はその編集者の過去の編集回数でのみ説明されることを表すモデルである．

しかし本研究では編集者の編集回数の正確な予測は必要ではなく，編集者の傾向，すなわち編集を続けるか止めるか，続ける場合の編集回数の規模が主な焦点である．また，2010年以降に参加した編集者も存在し，そのような編集者は単独で時系列モデルを構築するには情報が少なすぎるという問題がある．以上の理由から自己回帰モデルは適用しない．

そこで，本研究では自己回帰モデルと教師あり学習を用いて，自己教師あり学習 (Auto-Supervised Learning) アプローチを提案する．具体的な予測モデルを以下に示す．

$$y_q = f(y_1, y_2, \ldots, y_{(q-1)}, x_1, x_2, \ldots, x_m) + \epsilon_q \quad (2)$$

$x_1, x_2, \ldots, x_m$ は期間 $[1, q-1]$ における編集回数以外の m 個の特徴ベクトル，f は予測関数，ϵ_q は誤差である．つまり，過去の編集回数とそれ以外の編集者の特性を表す特徴を入力変数とし，ある予測関数を用いて未来の編集回数を出力する．期間 $[2, q]$ の編集情報をテストデータとして学習器に入力して $q+1$ の編集回数 p を予測する．これは以下の式で表される．

$$p = f(y_2, y_3, \ldots, y_q, u_1, u_2, \ldots, u_m) \quad (3)$$

ここで，$u_1, u_2, \ldots, u_m$ は期間 $[2, q]$ における編集回数以外の m 個の特徴ベクトルとする．このアルゴリズムを図 1 に示す．ここで，D はデータの全体集合を表す．自己教師あり学習法の特徴として，学習用サンプルをテストサンプルとして再利用できるという利点がある．

3.2 特徴空間

編集者の特性を表す可能性のある特徴ベクトルを独自に選択し，これを説明変数とする．特徴空間の概要を表 1 に示す．使用した特徴は，主に総編集数，総編集記事数，名前空間の編集率，編集文字数に関する特徴，編集記事属性，コメント，最後の期間の情報量，重み付き情報量等である．名前空間の編集率は，編集者の全ての編集文字数に対する各名前空間の編集文字数の割合として定義する．影響度について説明する．記事への影響度として，(4) 式で表される特徴を定義する．

$$i = \begin{cases} \dfrac{\Delta c}{c} & (\Delta c \geq 0) \\ \dfrac{|\Delta c|}{c - \Delta c} & (\Delta c < 0) \end{cases} \quad (4)$$

ここで，Δc は編集文字数，c は編集後の記事の文字数である．編集文字数から記事への影響度を定量化した特徴である．次に，図 2 は時間と編集者の数の関係を表す．図 2 から，2010 年からは約 2 倍に増加していることがわかる．さらに，最近の 5 カ月間の編集数をそのまま予測値として計算した場合，WMF の基準を 23.7%改善する．以上より，全ての編集者の属性は最近の編集情報に多いことがわかる．そこで，直近の期間に高い重み付けを行う重み付け情報量の特徴を定義する．最後の期間の情報量として編集文字数，影響度，編集回数増加量，コメント文字数を使用する．重み付け情報量にはさらに編集回数，編集記事数も使用する．(5) 式で表される．

$$S_w = \sum_{t=1}^{p-1} \frac{100}{p-t} I_t \quad (5)$$

リバートとは編集の差し戻しであり，ある編集者の行った編集が別の編集者によって差し戻されることがある．リバートに関する特徴は，[Suh 09] で示された Wikipedia 新参者がリバートによってコミュニティを離れる可能性に考慮した特徴である．

カテゴリ，リダイレクト，タイトル等の記事属性に関する特徴は，直接的には編集回数に関係しないと考えられる．しかし，データマイニングにより隠れた関係性を発見するためにもこれらの特徴を使用する．

Algorithm 1 Predict the number of edits vector p

Input: D

Output: p

1: Split D into $D_1, D_2, \ldots, D_q$
2: $D_{train} \Leftarrow \cup_{t=1}^{q-1} D_t$
3: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ from D_{train}
4: $\boldsymbol{y}_q$ from D_q
5: $Train\ f(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{(q-1)}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m) \Rightarrow y_q$
6: $D_{test} \Leftarrow \cup_{t=2}^{q} D_t$
7: $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m$ from D_{test}
8: $\boldsymbol{p} \Leftarrow f(\boldsymbol{y}_2, \boldsymbol{y}_3, \ldots, \boldsymbol{y}_q, \boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m)$

図 1: 自己教師あり学習アルゴリズム

表 1: 特徴空間の概要

番号	特徴
(1-2)	総編集回数と総編集記事数
(3-8)	名前空間 1-5 の編集率
(9-10)	編集文字数の平均と文字追加率
(11)	影響度の合計
(12-13)	自身でリバートした回数と他人にリバートされた回数
(14-15)	手続き型カテゴリと評価型カテゴリの記事数
(16-17)	リダイレクト記事編集回数と関連記事付き記事数
(18-19)	編集回数増加率と減少率
(20-21)	編集回数増加量の最大値と減少量の最大値
(22)	記事のタイトル文字数の平均
(23-25)	コメント文字数の平均，最大，最小値
(26-28)	編集期間数，活動的な編集期間数，編集時間の長さ
(29-32)	最後の期間の情報量
(33-39)	重み付け情報量
(40-)	期間ごとの編集回数 (y_t)

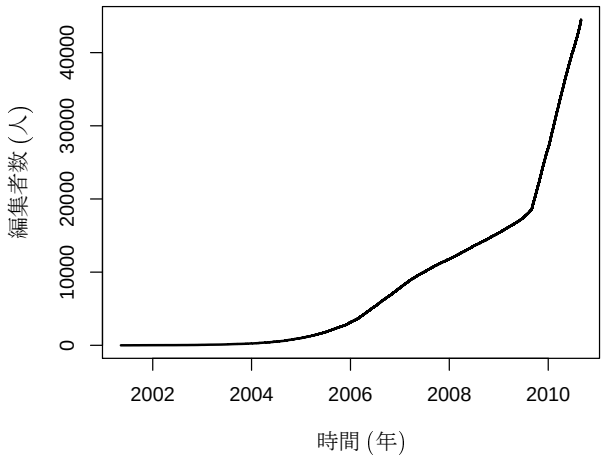

図 2: 時間と編集者数の関係

3.3 部分最小二乗回帰 (PLSR)

独自に選択した変数間には高い相関関係が見られた．多重共線性が存在する場合線形重回帰モデルの信頼性は低くなる．そこで多変量解析による回帰手法である部分最小二乗回帰法 (PLSR) を使用する．PLSR は計量化学分析の分野でよく用いられる手法である．変数集合から相関の低い潜在変数 T を抽出し，T を説明変数として線形回帰を行う．多重共線性問題を回避し，主成分回帰 (PCR) と比較してより目的変数との相関

表 2: 各処理の RMSLE 値

	RMSLE
無変換	2.297
対数変換	1.061

表 3: 各時間分割における RMSLE 値

	RMSLE
等分割	1.061
(B-1)	1.162
(B-2)	0.998

が強い変数を使用するため，予測問題に対しては有効な手法である．潜在変数抽出には [Dayal 97] で示されたカーネルアルゴリズムを使用する．クロスバリデーションによる予測誤差に基づいて，Wold の R 基準 [Wold 78] で最適成分数を決定する．

3.4 チューニング

データの特徴に関して，編集者ごとの総編集数を四分位分析した所，非常に大きい偏りが存在した．そこで，編集回数に関する最適な変数を選択するために，対数スケールを取った場合とのクロスバリデーションによる誤差を比較する．誤差は後述する RMSLE 値を用いる．結果を表 2 に示す．ただし，時間は等分割とする．表 2 から，対数スケールを取ると精度が良いことが分かる．そのため，編集回数に関する特徴は対数スケールを取る．

次に，より精度が高い時間分割法を探索する．時間分割法は以下のように定義する．

Step1 T_p を 2010 年 9 月 1 日，T_0 を 2001 年 1 月 1 日とする．

Step2 $T_{(p-1)} = T_p - d$ とする．ここで d は 153 日を表す定数とする．

Step3 以下の (B-1) と (B-2) を比較する．

(B-1) $T_{(p-n)} = T_{(p-1)} - (T_{(p-1)} - T_0)(1 - \frac{1}{b^n})$ とする．ここで，$n = 1, 2, \ldots, 15$，である．

(B-2) $T_{(p-n)} = T_{(p-1)} - b^n$ とする．ここで，$n = -4, -3, \ldots, 12$ である．

上記はトレーニングセットにおける時間であり，テストセットでは 5ヶ月時間を進めるものとする．(B-1) は最近の期間を長くとり，昔の期間を短くとる時間分割法である．(B-2) は逆に最近の期間を短くとり，昔の期間を長くとる時間分割法である．各場合のクロスバリデーションによる RMSLE 値を表 3 に示す．表 3 から，(B-2) の時に最も精度が良いことがわかる．そのためこの時間分割法を用いる．

4. 実験

実験の目的は，予測の精度を関連研究の手法と比較することである．データに含まれる編集者の総数は 44514 人であるが，期間 $[1, p-1]$ の間に編集を行った者は 33839 人のみであった．このためトレーニングサンプル数はテストサンプル数より少なくなる．実装において，分析と回帰予測にはオープンソースの統計解析ソフトである R を使用したが，特徴空間の算出には java を使用した．データベースを使用したデータの結合処理は java による並列処理プログラミングで高速に計算を行った．部分最小二乗回帰は R のパッケージである pls[Mevik 07] を使用した．

4.1 データセット

データには WMF より提供されたもののみを用いる． これは編集者 ID や編集記事 ID，編集時間，名前空間，編集文字数等の過去の編集情報を含む．編集時間情報を基に，データを分割して自己教師あり学習を実行する．予測回帰モデルを評価する指標として (6) 式で表される RMSLE 値を使用する．

$$\epsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(1+p_i) - \log(1+a_i))^2} \qquad (6)$$

ここで，n は編集者の総数，p_i と a_i は編集者 i の予測値と実測値を表す．これは，個々の編集者の編集数の正確な予測よりも編集活動の継続と停止，編集回数の規模がより重要な情報であると考えられるためである．

4.2 結果

テストセットに適用した際の RMSLE 値を表 4 に示す．改善率は WMF の基準値に対する誤差の改善率を表す．提案手法は関連研究と比較して予測精度が低いが，WMF の基準値を 40.5 %改善した．

また，使用した特徴について考察するために [Herring 11] のモデル同様ランダムフォレスト重要度 [Strobl 08] を算出する．重要度の上位 20 を表 5 に示す．表 5 は重要度が最大の特徴を 1 とした時の各特徴の重要度の値を表している．

5. 関連研究

編集行動の予測に関しての同時期に作成された予測モデルについて説明する．J.Moser は後述する予測評価値 RMSLE を最小にする最適定数を用いて予測を行った*3．つまり，全ての編集者が将来同じ回数の編集を行うと仮定した時，RMSLE 値が最小となる編集回数を探索した．これを WMF による基準値とする．さらに J.Moser は，予測すべき期間の直前の 5ヶ月間における編集回数とした場合，WMF の基準を 23.7%改善することを示した．

[Herring 11] は WMF から提供されたデータの他に，独自に Wikipedia から抽出したデータを使用し，高い精度の予測を行った．予測方法の特徴は，編集者の編集時系列に含まれる多くの性質の異なる特徴が将来の編集行動において重要な役割を果たすことを予想した上で，大量の特徴を学習アルゴリズムに投入していることである．[Zhang 11] は編集回数と時間に関する特徴のみを使用し予測を行った．予測期間の編集回数

表 4: テストセットに適用した際の RMSLE

手法	RMSLE	改善率
提案手法	0.881	40.5%
K.T.Herring	0.840	43.2%
D.Zhang	0.863	41.7%
WMF baseline	1.480	0%

*3 http://www.heritagehealthprize.com/c/hhp/forums/t/661/the-optimized-constant-value-benchmark/4330#post4330

表 5: 各特徴の重要度

特徴	重要度
重み付け情報量 (編集回数)	1.000
総編集回数	0.683
編集時間の長さ	0.664
重み付け情報量 (編集記事数)	0.550
重み付け情報量 (編集回数減少数)	0.508
活動的な編集期間数	0.482
重み付け情報量 (コメント長)	0.483
編集回数 (y_4)	0.444
総編集記事数	0.437
重み付け情報量 (編集回数増加数)	0.416
重み付け情報量 (影響度)	0.349
編集回数増加量の最大値	0.294
影響度の合計	0.288
編集回数 (y_3)	0.285
編集回数減少量の最大値	0.282
編集回数 (y_2)	0.271
編集期間数	0.260
編集回数減少率	0.220
重み付け情報量 (文字数)	0.156
編集回数 (y_5)	0.146

と学習期間の予測回数の平均 d を算出し，勾配ブースティングによる予測値に d を加えて補正するアルゴリズムを提案している．これは，編集者の直近の動向を予測に反映させることで，より編集者の最近の状態を表すモデルを構築するアルゴリズムであると言える．

6. 考察

予測精度について，提案手法は関連研究より低い．[Herring 11] のモデルで使用された学習サンプル数は約 250 万であるのに対して，提案手法で使用した学習サンプル数は 33839 のみであるためと考えられる．[Herring 11] のモデルはデータセットの偏りによる影響が少ないモデルである．提案手法は学習サンプルとテストサンプルを同一のデータから得ているため，データ量に制限がある状態で予測を行う場合は提案手法は有効であると言える．

提案手法は [Herring 11] のモデルと同様，無関係と考えられる多くの特徴を投入することで，隠れた編集回数との関係性を発見することができると考えて設計した．一方，[Zhang 11] は編集回数，記事数，編集期間のみのデータを特徴として使用して高い予測精度を得ている．これらの特徴は提案手法も含み，サンプル数も提案手法と差がほとんど無い．以上より，編集者の直近の動向を予測に反映させる補正アルゴリズムが，提案手法よりも予測精度を高めると考えられる．

次に，使用した特徴について考察する．[Herring 11] は編集回数に関する特徴以外にも，編集者の特性を表す特徴を多く使用した．しかし結果は編集期間と編集回数に関する特徴が予測モデルの大部分を占め，予測に有効だと予想した投稿ブロック等の特徴は低い重要度を示した．[Zhang 11] が使用した特徴は既に述べた．本研究も同様に，隠れた関係性が存在すると考えた編集記事属性などの特徴は低い重要度を示した．特に，[Suh 09] で示されたリバートによる編集数の減少が予測に影響すると考えたが，リバートに関する特徴も低い重要度であった．高い重要度を示したのは，関連研究同様，編集回数，編集記事数，編集期間に関する特徴である．さらに，独自に定義した重み付け情報量，記事への影響度に関する特徴も高い重要度を示した．特に，重み付け情報量の多くが高い重要度を示した．このことから，時間的情報価値を考慮した特徴は予測に有効であると言える．

以上をまとめると，未来の編集回数を予測するのに使用すべき特徴は，過去の編集回数，編集記事数，過去から現在までの編集時間から得られる情報である．時間的価値を考慮したその他の特徴も予測精度を高める．しかし，多くの特徴を使用するよりも編集者の直近の動向を予測に反映させる補正アルゴリズムを使用することがより有効であると考えられる．

7. 結論

本研究では過去の編集情報から，編集者の編集行動を予測する手法の提案を行った．予測には自己教師あり学習と独自に選択した特徴空間を用いた．また予測には最小二乗回帰を使用した．予測精度は関連研究よりも低いという結果が得られた．実験による特徴評価の結果，過去の編集回数，編集記事数，過去から現在までの編集期間から得られる情報，時間的価値を考慮した特徴が予測精度を高める特徴であった．

隠れた関係性が考えられる多くの特徴を使用する提案手法よりも，編集者の直近の動向を予測に反映させる補正アルゴリズムを使用することがより有効であると示された．

参考文献

[Suh 09] B.Suh, G.Convertino, E. H.Chi, and P.Pirolli: The singularity is not near Slowing growth of Wikipedia(2009) Proceedings of the 2009 International Symposium on Wikis (WikiSym), Orlando, FL, USA.

[Herring 11] K.T. Herring ,"Wikipedia Participation Challenge Solution" `http://meta.wikimedia.org/wiki/Research:Wiki_Participation_Challenge_Ernest_Shackleton` (last accessed at 2011-12-29)

[Zhang 11] D. Zhang ,"Wikipedia Edit Number Prediction based on Temporal Dynamics Only" `http://arxiv.org/abs/1110.5051` (last accessed at 2011-12-29)

[Nakata 10] 中田　和宏, 西山　裕之, 大和田　勇人,"Swing による視覚化を利用した肝細胞癌再発予測支援" , FIT2010(第 9 回情報科学技術フォーラム)

[Crone 10] Sven F. Crone, Nikolaos Kourentzes ,"Feature selection for time series prediction ? A combined filter and wrapper approach for neural networks", Neurocomputing 73 (2010) 1923-1936

[Dayal 97] Bhupinder.S.Dayal,John.F.MacGregor "Improved PLS Algorithms" JOURNAL OF CHEMOMETRICS,VOL.11,73-85(1997)

[Wold 78] Wold, S. "Cross-validation estimation of the number of components in factor and principal component analysis". Technometrics 24, 397-405.(1978)

[Mevik 07] Bjorn-Helge Mevik, Ron Wehrens,"The pls Package: Principal Component and Partial Least Squares Regression in R," Journal of Statistical Software,2007, Volume 18, Issue 2.1-24

[Strobl 08] Carolin Strobl, Anne Boulesteix, Thomas Kneib, Thomas Augustin, Achim Zeileis,"Conditional variable importance for random forests", BMC Bioinformatics,9:307 (2008)