4M-1

過去の編集情報を用いたWikipedia利用者の編集行動に関する予測

吉田　裕†　　大和田　勇人†
†東京理科大学 理工学部　経営工学科

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．それ故に,参加者の将来の編集行動を決定する要因を定量的に理解することはコミュニティの規模と多様性の成長を維持するためにも重要である．

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．同時に,,編集者の編集行動に強く関わる特徴を定量的に評価することを目的とする．

2　関連研究

B.SuhらはWikipediaの編集数と編集参加者数両方の1カ月当たりの成長率の減少について研究を行った．それに加えてWikipedia内の論争と支配，抵抗関係の存在によって，新しい寄与行動の機会が制限される結果が示唆されている．これも編集者行動の予測における重要な特徴となり得る[1].

編集行動の予測に関しての同時期に行われた研究は他にも存在する．D.Zhangは編集者の特徴に編集時間データのみを使用し，機械学習の一種である勾配ブースティングを用いて高い精度の予測を行った[2]．本研究と並行して行われ，同問題を扱うために類似点が多いが，本研究は実験に基づく特徴評価を行っている点でこれとは異なる．

Wikipedia Edit Number Prediction from the Past Edit Record
†Yutaka YOSHIDA　‡Hayato OHWADA
†Department of Industrial Administration, Faculty Of Science and Technology,Tokyo University of Science

3　データセット

データにはWMFより提供された英語版Wikipediaにおける2001年1月1日から2010年9月1日までの編集記録情報を用いる．予測する編集数は各編集者の2010年9月1日から2011年2月1日の5カ月間に行われたものとし，これを予測する回帰モデルを構築する．データは編集者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．さらに，個別の記事データを参照して，編集された記事の情報も関連付けることができる．予測回帰モデルを評価する指標として(1)式で表されるRMSLE値を使用する．

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

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　提案手法

4.1　自己教師あり学習 (Auto-Supervised Learning)

基本的なアプローチは，自己教師あり学習による連続値データの回帰予測である．これは時系列解析におけるARモデルを参考にし，機械学習による非線形回帰法を取り入れた．データDはある時間間隔$t(t = 1, 2, \ldots, p)$によって$D_1, \ldots, D_p$分割される．データD_tの編集数を$\boldsymbol{y}_t$とする．$\cup_{t=1}^{p-1} D_t$に含まれる個数mの特徴$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$と過去の編集回数$\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{(p-1)}$を説明変数とし，$\boldsymbol{y}_p$を目的変数とするトレーニングセットから回帰モデルを構築する．具体的なモデル式は(2)式を用いる．

$$y_p = f(y_1, y_2, \ldots, y_{(p-1)}, x_1, x_2, \ldots, x_m) + \epsilon_p (2)$$

fは予測関数，ϵ_pは誤差である．予測関数にはランダムフォレスト法(RF)と部分最小二乗回帰(PLSR)を使用した．これは，特徴評価が可能な点と，多重共線性回避のためである．交差検定による予測誤差を損失関数とし，損失関数を最小にするモデルを探索する．その後，$\cup_{t=2}^{p} D_t$に含まれるm個の特徴$\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$と$\boldsymbol{y}_2, \ldots, \boldsymbol{y}_p$を説明変数としたテストセットを最適モデルに適用し，将来の編集数の予測を行う．このアルゴリズムの概要をAlgorithm1に示す．

Algorithm 1 Calculate $RMSLE$

Input: D, A

Output: ϵ

1: Split D into $D_1, D_2, \ldots, D_p$
2: $D_{train} \Leftarrow \cup_{t=1}^{p-1} D_t$
3: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ from D_{train}
4: $\boldsymbol{y}_p$ from D_p
5: $min\ L\{f : \boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{(p-1)}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m \Rightarrow y_p\}$
6: $D_{test} \Leftarrow \cup_{t=2}^{p} D_t$
7: $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m$ from D_{test}
8: $\boldsymbol{a}$ from A
9: $\boldsymbol{p} \Leftarrow f(\boldsymbol{y}_2, \boldsymbol{y}_3, \ldots, \boldsymbol{y}_p, \boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m)$
10: $\epsilon \Leftarrow \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(1+p_i) - \log(1+a_i))^2}$

4.2 特徴空間

編集者は2010年からは約2倍に増加している．さらに，WMFが用意したベンチマーク結果によれば，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RMSLEは1.129となる．以上より，全ての編集者を特徴づける情報は最近の編集情報に存在すると言える．これを考慮して，主観的に特徴空間を選択した．特徴空間には編集文字数や編集期間の長さ，記事への影響度といった特徴も使用している．

5 実験

実験では交差検定によるRMSLE値を比較する．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(最小1，最大334173)．このためそのままの場合(1)と対数変換する場合(2)の精度を比較した．表1に示す．結果から編集回数に関する特徴は対数変換すれば精度が高くなることが分かる．

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．一つ目は時間を154日間隔で等分割する．それ以外を以下に示す．各場合の交差検定によるRMSLE値を表2に示す．結果から，(ii)の時精度が高いことが分かる．最後にテストセットに適用した予測精度を表3に示す．

(1) T_p を2010年8月31日，T_0 を2001年1月1日とする．

表1: 各処理のRMSLE値

	RF	PLSR
(1)	1.435	2.297
(2)	1.007	1.061

表2: 各時間分割におけるRMSLE値

	RF	PLSR
(i)	1.143	1.162
(ii)	0.954	0.998

表3: テストセットに適用した際のRMSLE

手法	RMSLE
RF	0.990
PLSR	0.881

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

(3) $b = 2$ として，以下の(i)と(ii)を比較する．

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

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

6 評価と考察

編集回数については対数をとり，時間間隔は最近の期間を短く，昔の期間を長く取れば精度が高い．評価値はRFよりもPLSRの方が精度が高かった．この結果はWMFが示した予測モデルを40.5%改善する．

7 まとめ

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低いが，実験に基づいた特徴評価ができたと言える．

参考文献

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

[2] D.Zhang, “ Wikipedia Edit Number Prediction Based on Temporal Dynamics Only ”,2011