評判抽出のためのブログ分類手法の比較検討

Comparing Weblog Classification Methods for Extracting Reviews

森田 悠基[*1] 松井 藤五郎[*2] 大和田 勇人[*2]
Yuki Morita Tohgoroh Matsui Hayato Ohwada

[*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]東京理科大学 理工学部 経営工学科
Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science

In this paper, we compared two weblog classification methods for extracting reviews. So far, we proposed the system for weblog search to get some reviews of products or services. And we proposed two approaches for that. First approach is a way to classify weblog articles into personal ones and non-personal ones. Second approach is a way to evaluate how much review information articles have and to rank these articles. We thought to improve the first approach. And we compared classification methods. As a result, combination of Naive Bayes and EM Algorithm performed better than combination of SVM and EM Algorithm. So we selected former method.

1. はじめに

本研究の目的はインターネット上のブログ空間から評判ブログのみを抽出し、さらに集めた評判ブログを容易に検索できるインターフェースをもった、評判ブログ自動収集・検索システムの構築である。ブログには個人が日記の用途で書いた主観性の高いブログ以外にもメモ代わりに使われるブログや、個人ユーザ以外による広告ブログ、スパムブログといったブログが存在する。評判ブログの検索システムを構築する上で、個人による主観を含むブログ以外のブログをうまく取り除く必要がある。そこでまず本研究では以下のようにブログ空間を仮定する。

- ブログ空間は【個人ブログ】と【非個人ブログ】の２つに分割可能である。
- ブログ空間は【意見を含むブログ】と【事実のみを伝えるブログ】の２つに分割可能である。

この仮定のもと本研究では、【個人ブログかつ意見を含むブログ】が求める評判ブログと一致すると考える。

また、本研究における評判ブログとは『個人ユーザによる、ある製品やサービスに対する主観的な意見を含むブログ』と定義する。

これまでの研究 [4, 5] では【個人ブログ】と【非個人ブログ】の分類のために、機械学習の 1 種である Naive Bayes と EM アルゴリズムを組み合わせた手法による分類を用いた。さらに【意見を含むブログ】を抽出するために、新聞記事を訓練データに用いてブログの意見性スコアリングを行い、ブログの意見量を数値化し、ランキングする手法を行ってきた。このような２段階手法によって評判ブログをランキング上位に抽出することができた。

連絡先: 森田 悠基, 東京理科大学大学院 理工学研究科 経営工学専攻 大和田研究室, 千葉県野田市山崎 2641, 04(7124)1501, morita @ ohwada-lab.net
2009/04 以降、松井藤五郎の所属は [とうごろう機械学習研究所] に変更されました.

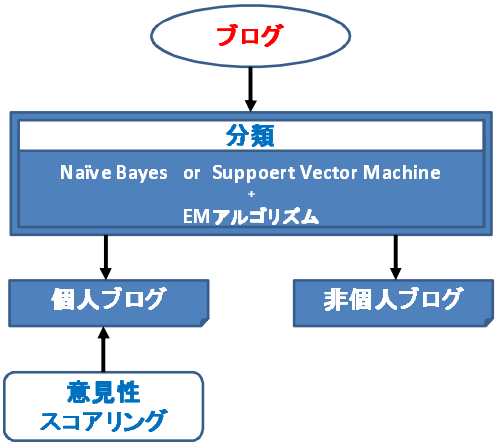

図 1: システムの分類手法

本論文ではこれまでの個人・非個人ブログへの分類の精度を改善していくために、分類手法を検討することを考える。まず、個人・非個人ブログへの分類手法について説明し、その分類手法を検討していくための比較実験を行っていく。個人・非個人ブログへの分類のための分類アルゴリズムとして、本研究では Naive Bayes と EM アルゴリズムを組み合わせた手法、SVM と EM アルゴリズムを組み合わせた手法の 2 つを検討する。(図 1)

また、本研究では毎日増えるブログを日々収集・分類しつづけることでデータベースを増やしていくことを考える。

2. 関連研究

Xiaochuan Ni ら [1] はブログ空間が informative ブログと affective ブログの 2 つに分割で可能であるとし、3 つの分類アルゴリズム (Naive Bayes, SVM, Rocchio) と 2 つの特徴語選択手法 (Information Gain,CHI-square statistics) による分類比較実

験を行っている。ここで informative ブログとは趣味や専門的な知識、ビジネス関連トピックに関するブログ記事のことを指し、affective ブログとは個人が感じたことなどを日記として書き表したものを指している。この分類の結果では SVM と Information Gain による組み合わせのアルゴリズムが分類に適していた。Xiaochuan らはこの分類の高い精度から、ブログ空間が確かに informative ブログと affective ブログに分類が可能であったとしている。

Xiaochuan Ni らの研究は中国のブログ空間に対して行っており、日本語のブログ空間のように活用形を持っている言語に対しても同じ精度がでるとは限らない。また information-affectiveness 分類と本研究の個人・非個人分類は少し異なっている。Xiaochuan らの分類はいくつかの応用先を考えて汎用的に行っているのに対し、本研究の目的は評判ブログの抽出と検索であるため、個人・非個人分類のほうが適していると考える。

3. 手法

本研究ではブログを個人ブログと非個人ブログに分類するために Naive Bayes と EM の組み合わせと SVM と EM の組み合わせによる分類を比較する。まず Naive Bayes と SVM について説明し、次にそれらを EM と組み合わせる手法を説明する。

3.1 Naive Bayes

Naive Bayes は与えられた訓練データ（既存データベースのラベル付きブログ）をもとに分類器を作成し、テストデータ（新規記事）の各クラス $V_j (j \in \{$ 個人・非個人 $\})$ に対する確率を推定する。あるインスタンス（ここではブログの特徴語）$\langle a_1, \cdots, a_n \rangle$ を含む文書 d（ブログ記事）が与えられたとき、一番もっともらしい文書のクラス V_{MAP} は

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, \cdots, a_n) \tag{1}$$

と表すことができる。この式にベイズの定理を用い、さらに各インスタンスが独立であると仮定すると文書のクラス v_{NB} を求める式は、

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i | v_j) \tag{2}$$

となる。このアルゴリズムを用いてトレーニングデータから分類器を作成しテストデータのクラス推定を行う。本研究では、データマイニング・ツール WEKA[*1] に含まれている Complement Naive Bayes[2] を用いる。

3.2 SVM

Support Vector Machine(SVM) はテキスト分類に対し高い精度をもつ教師あり学習アルゴリズムである。その一番シンプルな形である線形 SVM は正事例と負事例のマージンを最大化することで、正事例と負事例を 2 つに分割可能な超平面を発見する。本研究では、WEKA に組み込まれている LIBSVM[*2] というライブラリを用いて、C-support vector classification（C-SVC）を用いる。

3.3 EM アルゴリズムによるラベルの決定と分類器の更新

EM アルゴリズムは分類アルゴリズム (Naive Bayes, SVM) と組み合わせることで次のように適用される。

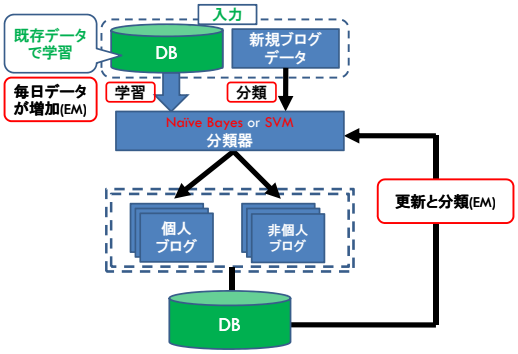

図 2: EM アルゴリズム

1. ラベルつきデータとラベルなしデータ（テストデータ）を入力。
2. 分類アルゴリズムにより分類器の作成および、ラベルなしデータのラベル尤度の測定。
3. 2 で得た結果から一時的なラベルをラベルなしデータに付与。
4. 3〜4 を繰り返し行い分類器が更新されなくなったらラベルを決定し終了。

E ステップによってラベルのついていないブログ記事の各ラベルに対する尤度を計算し、M ステップにて各ラベルを一時的に付与することで、ラベルの付いていないブログ記事データが一時的にラベルつきデータとなり、分類器を更新する。これによりもう一度 E ステップを計算するときに違う結果がでるため、E ステップと M ステップを繰り返すことで、何度もラベル付けを修正し、ラベルが変わらなくなるまで EM アルゴリズムによる計算を行う。

3.4 Naive Bayes, SVM と EM アルゴリズムによるデータの更新

これらの手法を組み合わせて適用する場合、まずトレーニングデータ (ラベルつきデータ) を用意し、分類器を作成する。そして前日に書かれたブログ (ラベルなしデータ) を事前に収集しておき、作成した分類器を用いてブログを分類し、一時的なラベルをブログに付与する。そして、再度一時的なラベルを付与されたブログから分類器を更新し、再分類を行っていく。終了条件は再分類をおこなっても分類器が更新されなくなったときとなる。ただし、再分類をおこない過ぎると過学習となる可能性もあるため、本研究では再分類の回数に制限を加えることも考える。

実際には、事前にトレーニングデータとしてブログをいくつか人手により分類し、トレーニングデータを作成する。そして 1 日目の分類では、このトレーニングデータから分類器を作成し、その日のブログを分類しデータベース化する。次に 2 日目以降は、それまでに作成したデータベースをトレーニングデータとして利用していき、分類を繰り返していくことで、日々増えていくブログに対応することが可能となる。本手法では、毎日、それまでにデータベース化してきたブログを次の日のトレーニングデータとして使うという意味で EM アルゴリズムを用いる (再分類のたびにその日のブログが増えていく) と共に、1 日の分類においても EM アルゴリズムを用いて、複数回の分類を行っている。(図 2)

∗1 http://www.cs.waikato.ac.nz/ml/weka/
∗2 http://www.csie.ntu.edu.tw/ cjlin/libsvm/

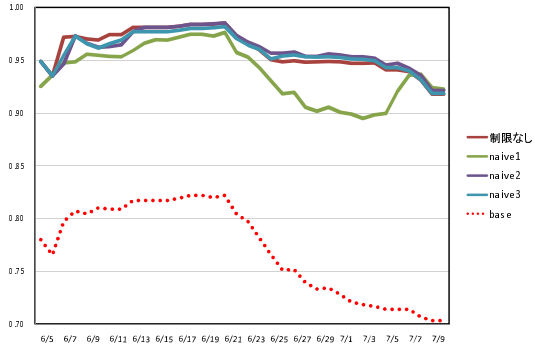

図 3: Naive Bayes/EM (Precision)

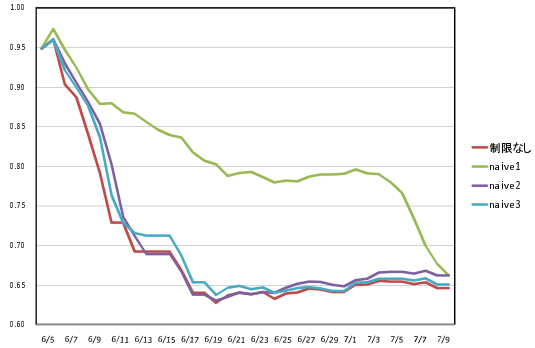

図 5: Naive Bayes/EM (Recall)

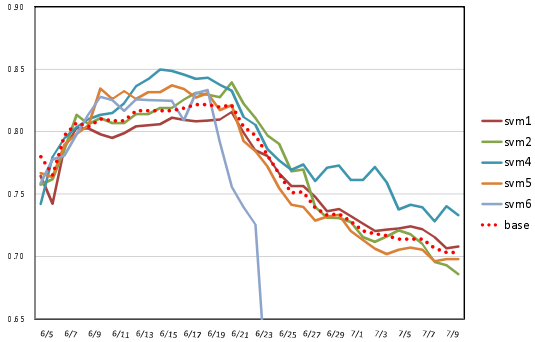

図 4: SVM/EM(Precision)

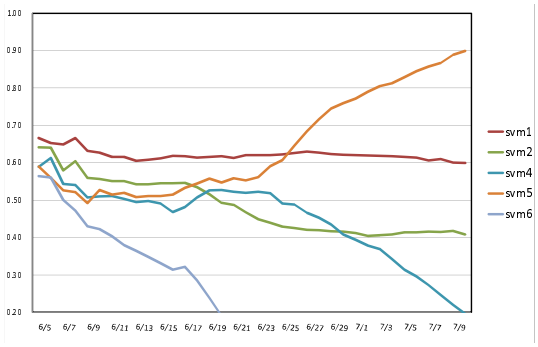

図 6: SVM/EM(Recall)

4. 実験

本章では、Naive Bayes/EM の組み合わせによる手法と SVM/EM の組み合わせによる手法を比較し、本研究におけるブログの個人・非個人分類に対しての有効性を確認する。

4.1 実験準備

本実験では、TechnoratiAPI を用いて検索キーワード「ixy」に関するブログを本システムのブログ収集手法によって取得した。取得したブログは 2008/06/05-2008/07/09 に集めたブログ 1254 件 (29 件をトレーニングデータとした) である。収集したブログには評価のために人手によって正解ラベルを付与した。そのうち個人ブログが 882 件、非個人ブログが 372 件である。

評価では *Precision,Recall* を用いる。*Precision* とは個人ブログとラベル判定したブログのうち、正しく個人ブログと判定できている割合」であり、*Recall* とは正しい個人ブログのうち、分類によって個人ブログのラベル判定をされたブログの割合」である。本研究では、評判ブログの検索を行った際に、実際に表示されるブログのなかに非評判ブログが混ざらないことを目的とするため、*Precision* を重視する。

4.2 実験結果および考察

Naive Bayes/EM の組み合わせ、SVM/EM の組み合わせによる実験の結果のうち *Precision* について図 3,4 に示す。*Recall* についても、図 5,6 に示す。naive1-3,svm1-6 とは 1 日での EM アルゴリズムによる再分類の回数を制限することで過学習を抑えた場合である。制限なしとは、そのような制限を設けずに実行した場合である。naive1 とは毎日 Naive Bayes による分類を 1 回だけ行ったことを示している。ただし、SVM/EM において制限なしで実行した場合、数日間データ更新していくうちに全ラベルが個人ラベルまたは非個人ラベルになってしまったために表示していない。また、base とは分類を行わず、収集したブログ群における個人ブログの割合を示す。この値を超えない場合、従来のブログ検索エンジンによる検索のほうが検索結果において個人ブログが表示される割合が高くなる。ただし、本研究では検索結果のランキング手法において意見性スコアリングによって評価値をつけているため、検索上位に個人ブログがうまく表示される可能性もある。ただし今回はそのような実験は行っていない。

実験の結果、Naive Bayes/EM による手法は *Precision* において常にほぼ 9 割を超えていることがわかる。これは base に比べ 1 割以上も高い値を保ち続けている。このため、実際に検索した結果においても従来の検索エンジンに比べ、評判ブログが表示される確率が高くなるといえる。ただし *Recall* においては、急激に値が下がっており、naive1 のように EM アルゴリズムを制限した場合のほうが過学習を抑えられる可能性があることがわかる。初期では急激に *Recall* の値が下がってしまっているが、6/17 頃を境にして *Recall* の値は安定しており、一定の値を保てるということが予想されるだろう。

次に SVM/EM による手法についても見ていく。SVM/EM の手法では *Precision* が大きく変動していまっていることがわかる。また、svm4,svm6 を除いて、ほぼ base と同じ値をとっていることがわかる。svm4 は初期値は低いものの、途中からは base よりも高い値をとっている。svm6 においては、6/17 頃から急激に値が下がってしまい、最終的には値が 0 になってしまった。これは、SVM がマージン最大化を行う上で、マージン周辺のデータのみを扱うため、再学習の度にデータを分割するための直線が一方のラベル側へと徐々に寄っていってしまっ

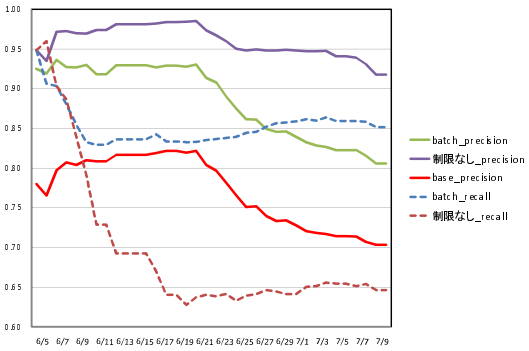

図 7: Naive Bayes によるバッチ処理分類との比較

たためだと考えられる。SVM においては 1 日における再学習回数を制限しないと、うまく学習ができなかった。*Recall* についても見ていくと、1 日 1 度しか分類を行わない svm1 の場合が一番安定した値をとりつづけていることがわかる。全体的には、学習回数を重ねるごとに値が悪くなってしまう傾向があった。ただし、svm5 においてのみ途中から値が急上昇し、高い *Recall* 値をとることになった。これは学習の頻度が本実験データの場合においてうまくいっただけの可能性が高く、一概に svm5 の場合が高い値をとるというわけではないだろう。

これらの結果から、本研究手法のように、個人・非個人にブログを分類する場合においては Naive Bayes と EM アルゴリズムを組み合わせる手法のほうが SVM と EM アルゴリズムを組み合わせる手法に比べ、*Precision*,*Recall* の両方において良い結果となった。特に、Naive Bayes/EM における *Precision* は高い値を取ることができ、本研究の趣旨とも合致することがわかった。

次に、本手法のように EM アルゴリズムを用いずに最初に用意したトレーニングデータのみを用いて Naive Bayes を用いて毎日ブログ分類を行った場合 (初期トレーニングデータによるバッチ処理) と、Naive Bayes/EM の組み合わせによる手法を比較した場合について図 7 に示す。直線グラフが *Precision* を示し、点線グラフが *Recall* を示している。

図からわかるように、*Precision* では、本研究による手法が明らかに batch 処理に比べ安定して高い値を出しており、有効であると考えることができる。しかし *Recall* においては、6/9 頃までは本手法および batch 処理の両方が値を下げているが、それ以降では本手法のみが値を下げてしまっている。また、6/19 以降はグラフの形はともに安定している。これらの結果から、batch による手法のほうが *F-measure* の値が高くなると考えられ、base と比べて *Precision* も高い値をとっていることから、本研究の分類方法 (個人・非個人分類) では確かに Naive Bayes が有効であることがわかる。ただし、本研究では最終的に検索されるブログにおける個人ブログおよび評判ブログの割合が高いことを好むため、Naive Bayes/EM の組み合わせによる手法における高い *Precision* のほうが有効であると思われる。

5. おわりに

本論文では、評判ブログの自動収集・検索をおこなうシステムにおける 2 段階手法 ([個人・非個人ブログ分類による個人ブログ抽出] と [意見性スコアリングによるブログの意見量の数値化]) のうち、個人・非個人ブログ分類に適用する分類手法について、Naive Bayes/EM の組み合わせによる手法と SVM/EM の組み合わせによる手法との比較実験・検討を行った。

実験の結果、Naive Bayes/EM による手法は 9 割を超える *Precision* を安定して出すことができた。SVM/EM による手法は EM アルゴリズムによる再分類の回数を適切に制限することで従来のブログ検索に比べ有効であった。また、これら 2 つの手法を比較したところ、総合的に Naive Bayes/EM による手法が有効であることがわかった。さらに Naive Bayes 単体によるバッチ処理分類との比較も行った。その結果、*Precision* では、Naive Bayes/EM による手法が明らかに有効であったが、*Recall* ではバッチ処理分類がより有効的であった。ただし、本研究では検索結果として表示されるブログにおける評判ブログの割合を重視しているため、*Precision* の高い Naive Bayes/EM による手法が、本研究において有効であるとした。

今後は分類時の特徴語選択に関しても検討していきたい。Yang ら [3] らは、DF, Information Gain, CHI-square といった手法を検討し、特に DF による特徴語選択では計算コストに対する効果が実用的であるとしている。私たちの手法に対しても有効であるか、比較などを行いさらに精度を上げていくことを考えていく。

参考文献

[1] X. Ni, G.R. Xue, X. Ling, Y. Yu, and Q. Yang. Exploring in the weblog space by detecting informative and affective articles. *Proceedings of the 16th International World Wide Web Conference (WWW-2007)*, 2007.

[2] Jason D. M. Rennie, LawrenceShih, Jaime Teevan, and David R. Karger. Tackling the poor assumptions of naive bayes text classification. *Machine Learning (ICML-2003), Washington DC*, 2003.

[3] Y. Yang and J.O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pp. 412–420. MORGAN KAUFMANN PUBLISHERS, INC., 1997.

[4] 森田悠基, 松井藤五郎, 大和田勇人. Brevis: ブログにおける評判情報自動収集・検索システム. 人工知能学会全国大会 (第 22 回),1E1-03, 2008.

[5] 森田悠基, 松井藤五郎, 大和田勇人. ブログにおける評判情報自動収集・検索システム brevis の開発. 日本ソフトウェア科学会第 25 回大会, 7B-2, 2008.