1E1-3

BREVIS：ブログにおける評判情報自動収集・検索システム

BREVIS: Automatically Collect and Search System for Review Weblog

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 propose a system which automatically extract and search review information from Weblog. For extracting review information, we combine two different approaches. First approach is classifying Weblog articles into personal articles and non-personal articles. Second approach is scoring opinions included in articles and sorting ordered by the score. Then, we can find articles which are written by individuals and have much review information. We experimented and could sort review information at the top of our system's searching result list.

1. はじめに

2006 年 3 月末のブログ登録者数は 868 万人を超えた [3]。ブログはその特性上、最新の話題に関する記事や個人の主観的な意見が多いことから、これらの情報を効率的に収集分析できれば意思決定支援や企業リスク管理といったことに利用できる。しかしこれらの情報を人手で集めていてはコストがかかってしまうため、自動で意見を収集、分析する研究が大学や企業で行われている。

本研究も対象にブログを使用する。インターネットのブログ空間から評判ブログを自動抽出し、それらを一般ユーザが検索できる評判ブログの自動収集・検索システムの構築を目的とする。

ブログには個人が書くブログ記事以外にも広告ブログやスパムブログが存在し、個人の主観的な意見としての評判を抽出するためには、これらをうまく取り除く必要がある。本研究では、ブログ空間を以下のように仮定する。

- 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 種類に分けることができる。
- ブログ空間は【個人ブログ】と【非個人ブログ】に分けることができる。

この仮定をもとに我々は【意見を含むブログかつ個人ブログ】は本研究で求める評判ブログと一致すると考える。

ここで本研究における評判ブログとは『ある製品・サービスなどに対する主観的な意見を含むブログ』と定義する。

我々は構築した評判ブログの自動収集・検索システムの提案として、ブログの自動収集、ブログへの意見性スコアリングによるブログの意見性の数値化、個人・非個人ブログの分類を行う手法を提案する。

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, 千葉県野田市山崎 2641, 04(7124)1501, j7408633@ed.noda.tus.ac.jp

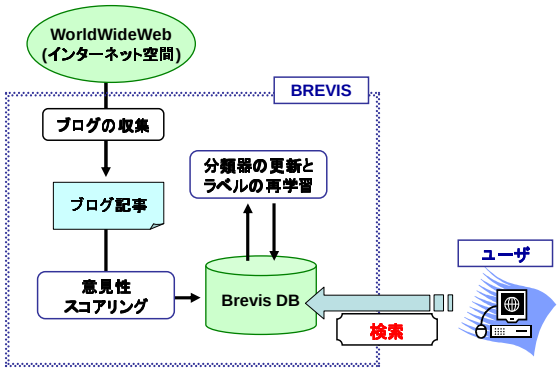

図 1: BREVIS の概要

2. 提案手法

本章では本研究で構築した検索システム BREVIS の評判ブログ収集の手法を提案する。2.1 節で BREVIS の特徴と流れについて述べ、2.2 以降で実際の手法について述べる。

2.1 BREVIS の特徴と流れ

本節ではまず BREVIS の特徴である評判ブログの抽出方法を説明する。本研究では【意見を含むブログかつ個人ブログ】を抽出するために 2 方向からのブログ分類を行う。1 つ目は、新聞記事を用いた特徴語辞書を使いブログの意見性スコアリングを行う手法である。これにより意見を含むブログを抽出することができる。2 つ目は NB/EM 分類によってブログを個人ブログと非個人ブログに分類する手法である。これによって個人ブログを抽出することができる。さらに、NB/EM 分類では過去の分類結果を次に引き継ぎ、新規ブログが追加される度に分類器の更新とブログラベルの再学習を行うことでインクリメンタルな分類を行う。

次に BREVIS の評判ブログの自動収集の流れを述べる。BREVIS の概要は図 1 のようになる。まずブログの収集を行い、取得したブログに意見性スコアリングを行い、ブログの意見性を

数値化する。つぎに NB/EM 分類によって新規ブログに個人ブログか非個人ブログかのラベルを付与する。 さらに新規ブログのラベルから分類器の更新とラベルの再学習を行う。

2.2 ブログの収集

ブログの収集には Technorati のキーワード検索 API を用いる。 これによってあるキーワードを含むブログを検索し、検索したブログのうち記事 URL、記事タイトル、サイトタイトルの重複のないブログのみをダウンロードしブログの取得を行う。本システムのブログ取得は毎日行い、1 回のブログ取得では前日 0-24 時の間に書かれたブログのみ取得する。

2.3 意見性スコアリング—新聞記事を用いた特徴語辞書によるスコアリング—

本システムでは意見を含むブログを抽出するため、新聞記事を用いた特徴語辞書を使いブログの意見性を数値化する。

本システムで抽出したい意見を含むような文章には意見を含む文章特有の概念を持った語（特徴語）が多く含まれていると仮定し、逆に意見を含まないような事実ばかりの文章には意見を含まない文章特有の特徴語が多く含まれると仮定する。

新聞記事は大衆に情報を正確に伝えるという特性上、一面記事や国際面などでは事実を文語体で表現する傾向にあるが、社説やインタビュー記事、投稿欄などは口語体で意見が書かれることが多いという特徴を持つ。本研究では毎日新聞 2006 年度版の記事のうち文語体で表現される傾向のある『国際記事』を非意見の集合体、口語体で表現される傾向のある『社説記事』と『みんなの広場という見出しの記事』を意見の集合体であるとみなし、これをトレーニングデータとして特徴語辞書を作成する。また新聞記事を用いることで、ドメインに偏らない信頼性のある大量のトレーニングデータを得ることができる. 特徴語のスコアは式 1 によって求められる。

$$score(w_i) = \frac{P_o(w_i) - P_F(w_i)}{P_o(w_i) + P_F(w_i) + k} \tag{1}$$

w_i はトレーニングデータにおける特徴語（形容詞、形容動詞、動詞）、$P_o(w_i)$ は特徴語 w_i が意見に含まれる確率、$P_F(w_i)$ は特徴語 w_i が非意見に含まれる確率をあらわす。定数 k は分母が 0 にならないようにするための定数である。これにより正の値が大きくなるほどその単語は意見文章に含まれ、負の値が大きくなるほど非意見文章に含まれると推測される。この特徴語辞書を用いてブログ記事 d の意見性スコア $Score(d)$ は式 2 によって計算される。

$$Score(d) = \sum_{w_i \in d} score(w_i) \tag{2}$$

2.4 分類器の更新とラベルの再学習—ナイーブベイズと EM アルゴリズム—

本システム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 NB/EM 分類 [2] を行う。 本節ではまずナイーブベイズ分類を説明し、次に EM アルゴリズムによるラベルの決定と分類器の更新を説明する。そして最後に毎日のシステム運用にこれを適用させる手法を説明する。

2.4.1 ナイーブベイズ分類

ナイーブベイズ法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

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, \cdots, a_n) \tag{3}$$

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、

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i | v_j) \tag{4}$$

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。

2.4.2 EM アルゴリズムによるラベルの決定と分類器の更新

EM アルゴリズムはナイーブベイズ法と組み合わせることで次のように適用される。

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. ナイーブベイズ法によってラベルつきデータからナイーブベイズ分類器 θ を計算する。
3. E ステップ：分類器を用いて、ラベルなしデータの各ラベルに対する確率を計算する。
4. M ステップ：E ステップで求めた結果からラベルなしデータに一時的なラベルを付与し、分類器 θ を更新する。
5. 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確率を計算し、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、ナイーブベイズ分類器を更新する。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。

2.4.3 新規記事の追加による分類器の更新

本手法の概要は図 2 のようになる。 本システムでは毎日取得する新規ブログに対し、既存のデータベースのブログラベルを利用してラベルを付与し、さらにそれによって変わったデータベースから分類器を更新し、ブログのラベルを再学習することで、毎日少しずつブログデータにラベルを付与しつつ、新しく取得したラベル付きブログデータを基にして、既存のデータベース内ブログデータのラベルを再学習することができるようになる。これは少数事例から一度に何日分もの大量のブログデータにラベルを付与するバッチ処理による不安定さに比べ、システムとして重要な安定性を得ることができる。

2.5 先行研究との違い

川口ら [5] は SVM 分類と意見文スコアリングによる 2 段階アプローチで評判ブログを抽出した。 川口らの手法の問題は SVM での分類時にトレーニングデータとして 100 件以上のデータを人手で用意する必要があったことである。本システムではこの問題を解決するため、少数事例からの分類に成功した杉田ら [4] の NB/EM 分類を採用した。これにより、本システムはシステム稼動前に 1 度だけトレーニングデータを用意するだけで以降は自動で評判ブログのみを収集する。

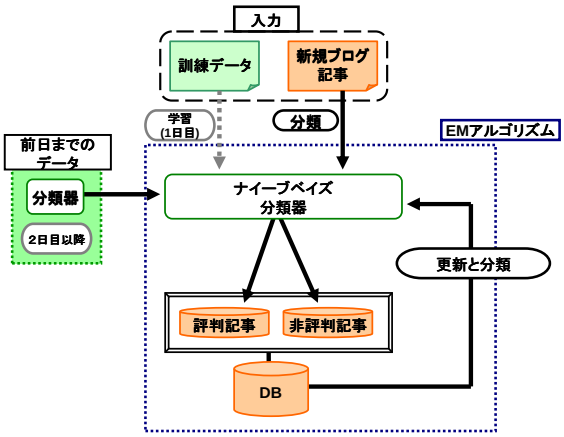

図 2: ナイーブベイズと EM アルゴリズムの組み合わせによる手法の概要

表 1: NB/EM 分類の結果

個人ブログ抽出	
Accuracy	0.734
Precision	0.524
Recall	0.917
F-measure	0.667

評判ブログ抽出	
Accuracy	0.677
Precision	0.413
Recall	0.897
F-measure	0.565

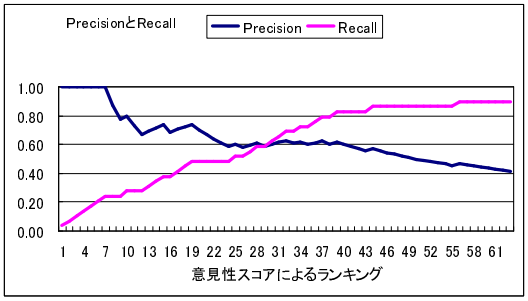

図 3: 分類結果

3. 評価実験

本章では作成した評判検索システムの有効性を確認するため次のような実験を行った。

3.1 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 2008/1/20,21 の連続する 2 日間のデータである。TechnoratiAPI による検索結果 194 件のうち提案手法により URL、ブログタイトル、サイトタイトルの重複した記事を除いた 126 件から文字コード変換に成功した 124 件 (評判ブログ 29 件：非評判ブログ 95 件) を用いた。新聞記事を用いた特徴語辞書の新聞データには毎日新聞 2006 年 [1] の国際・社説・見出しがみんなの広場である記事を使用し、特徴語として 5994 語を得た。また、NB/EM 分類における訓練データとして評判・非評判ブログ 5 件ずつを用意した。

3.2 評判ブログ抽出の評価

まず NB/EM 分類の結果を表 1 に示し、これに意見性スコアリングを行いスコア順にランキングし *Precision* と *Recall* を計算した結果を図 3 に示す。評価方法には *Accuracy, Precision, Recall, F-measure* を使った。*Accuracy, Precision, Recall* はそれぞれ正確度、精度、再現率を示し、*F-measure* は精度と再現率の総合的評価を行う指標である。

NB/EM 分類（表 1）により、124 件ブログのうち 63 件を個人ブログであると分類した。 これにより 3 件の個人ブログを非個人としてしまったが、58 件の非個人ブログを除外することに成功している。 またこれを評判ブログの分類と見た場合でも 3 件の評判ブログを非評判としてしまったが、58 件の非評判ブログの除外に成功している。

また、意見性スコアリングと組み合わせた結果（図 3）、抽出上位 10 件においては 80%の *Precision* を出している。また抽出上位 39 件で *Recall* が 80%を超えた。

4. 考察

提案手法による NB/EM 分類では *Recall* が高いことが特徴として挙げられる。これは個人（評判）ブログの見逃しがほとんどないことを示しており、ブログ取得時の個人（評判）ブログの数を維持したまま、非個人（評判）ブログの数を減らすことに成功したといえる。また、これに意見性スコアリングの結果を組み合わせランキングした結果抽出上位 10 件においては *Precision* が 80%あり、これは検索したときに上位 10 件中 8 件が評判ブログであることを示している。*Recall* に関しても抽出上位 39 件で 80%になっており、これは検索時に 39 件目までに 8 割の評判ブログがランク済みであることを示す。

ブログ取得時には 23.4% （124 件中 29 件）しかなかった評判ブログが分類後には 41.3% （63 件中 26 件）になり、さらに意見性スコアによってランキングすることで実際の検索時には抽出上位 50 件の中に 50% （25 件）の評判ブログが並ぶことになった。実際の検索ではユーザはページ遷移を何度も行うわけではなく、抽出上位に評判ブログが並ぶことで実際の *Precision* 以上の効果があると思われる。また *Precision* がこれ以上にならなかった理由の 1 つとして考えられるのは、スパムブログのなかには人がみても個人ブログと間違えそうなものもあり、これをうまく分類できていなかったことがあげられる。

5. 結論

本研究は評判ブログの自動収集・検索をおこなうシステム BREVIS を提案した。 評判ブログの抽出には新聞記事を用いた特徴語辞書によるスコアリングと NB/EM による分類の結果を組み合わせる 2 方向からの抽出アプローチでは、スコアリングによって意見を含むブログを抽出し、NB/EM 分類によって個人ブログの抽出を行い、これらを組み合わせることで評判ブログを抽出する手法を提案した。

実験では抽出上位 10 件において *Precision* が 80%、抽出上位 39 件にて *Recall* が 80%となった. これは従来のブログ検索エンジンによる検索に比べ、検索上位での評判ブログの割合が高く、検索時に評判ブログを得られる可能性が高くなった。このことから本システムが有効であるといえる。

参考文献

[1] 毎日新聞 CD-ROM (2006).

[2] Kamal Nigam, Andrew Kachites, McCallum Sebastian, and Thrun Tom Mitchell. Text classification from labled and unlabeled documents using EM. *Machine Learnig, Vol.39, No.2/3, pp.103-134,* 2000.

[3] 総務省 ブログおよび SNS の登録者, 2006.

[4] 杉田龍典. ナイーブベイズ法と EM アルゴリズムを用いたレビューと非レビューの自動分類. 東京理科大学理工学部経営工学科　卒業論文, 2007.

[5] 川口敏広, 松井藤五郎, 大和田勇人. 2 段階アプローチによる Weblog からの意見文抽出. 電子情報通信学会技術研究報告 Vol.106, No.473, 2007.