画像クラスタリング結果から得られる属性を用いたブログ記事の意見記事・非意見記事判定システムの設計と実装

佐々木 英文† 市川 哲彦‡ 田中 稔†

†山口大学大学院 理工学研究科 〒755-8611 山口県宇部市常盤台 2-16-1
‡山口大学 メディア基盤センター 〒755−8505 山口県宇部市南小串 1−1−1

E-mail: {k017vk, ichikay, tanakam}@yamaguchi-u.ac.jp

あらまし 本研究ではブログからの意見記事抽出システムの提案をする。本システムでは、既存システムで判定に用いられていた属性（ブログ中の単語、品詞の割合、アフィリエイトリンクの有無）に加え、ブログに含まれる画像から得られる導出属性を加えた判定を行っている。導出属性はブログに含まれる画像データの検索結果を分析し、含まれる画像の類似性を用いて得られたブログ記事クラスタから抽出される属性を用いた。導入した結果、5 割のデータで正確性、適合率、再現率のいずれかが向上した。また、8 割のデータで既存手法と同等以上であることがわかった。このことから、効率良く意見記事を発見する際に有用であると考えられる。また、本手法を利用した意見記事・非意見記事判定システムを試作した

キーワード ブログ, 意見記事抽出, 画像クラスタリング

A review and non-review classification system for Blog articles using the property derived by clustering the included images based on an image similarity measure

Hidefumi Sasaki† Yoshihiko Ichikawa‡ and Minoru Tanaka†

† Graduate School of Science and Engineering, Yamaguchi University Tokiwadai 2-16-1, Ube, Yamaguchi, 755−8611 Japan
‡ Media & Information Technology Center, Yamaguchi University Minamikogushi 1−1−1, Ube, Yamaguchi, 755−8505 Japan

E-mail: {k017vk, ichikay, tanakam}@yamaguchi-u.ac.jp

Abstract This paper proposes a system for selecting customer reviews from Blog articles using images in the articles. Previously know methods utilize text information such as appearance of words typically used in reviews, relative frequency ratios of nouns, verbs and adjectives, and existence of affiliate marketing links. A lot of Blog pages, however, include images of products, and non-review (or commercial) articles tend to use commercial photos of products. Hence, if we can discriminate commercial photos from personally taken photos, the information may be used as an additional factor in choosing reviews. The proposed system is a mediator between users and Blog search engines, and discriminates commercial photos as follows: (1) images in the search results are clustered based on an image similarity measure; and (2) photos in larger clusters are considered to be commercial. Our system uses this information as well as the textual properties to categorize search results into reviews and non-reviews. The prototype system utilizes SVMs (support vector machines) as a categorizer. Our experimental results indicate that either of accuracy, precision or recall improves in one half of the test cases, and that the results are no worse than the existing technique in 80% of the test cases. Therefore the proposed method is suitable for glancing quickly over reviews. We have implemented a prototype system using the proposed technique.

Keyword Blog, review extraction, image clustering

1. はじめに

本研究では、ブログ画像データの検索結果を分析し、含まれる画像の類似性を用いて得られたブログ記事クラスタから抽出される属性を用いた意見文判定処理手法の提案と実装を行う。

まず本研究の背景について述べる。World Wide Web（WWW）の普及に伴い、インターネットで利用可能な情報は、特定の管理者の元で生成され発信されていたものから、一般利用者自身によって生成されるものへと変化をした。これらのデータは HTML などの緩やかに構造化された文書形式をとり、また、メタデータが（基本的には）存在しないことから、内在する構造を抽出する研究がこれまでになされてきた[1][2][3]。これらが主にハイパーテキストの参照関係を用いた分析を行う手法である。

また、近年はブログ、Wiki、SNS といった情報発信ツールが普及するのに伴い、ユーザ生成コンテント（User-Generated Content, UGC）とよばれるさらに広い層の利用者による情報発信がなされるようになった。そのため、上述のデータマイニング手法は、UGC を分析する手法へと発展し、ブログページ間のトラックバック関係を利用することでコミュニティ分析や単語の出現傾向を調べることによる流行分析などがなされるようになった[4]。

これらの研究は大別すると、アンカーやトラックバックのように明示的に参照関係を利用するものと、テキストや画像などの内容から導出される関係を利用したものとに大別される。後者の例としては、ページ間に存在する単語の共起関係を用いてコミュニティ分析をするものや、あるいは、ブロガが利用しているブログカテゴリの類似性からブロガ間の類似関係を導きだし、それによってブロガコミュニティ分析をするというものがある[5][6]。

ブログデータの分析は市場分析という側面があるため、消費者の意見を反映した記事のみを抽出する意見記事抽出処理の研究もなされている。文献[7]の手法では、ブログ中の単語、品詞の割合、アフィリエイトリンクの有無など各ブログ記事から直接抽出可能な属性を因子として意見記事判定処理をしているが、ブログ中には他にも画像データ、トラックバック、コメントなどがあり、類似あるいは関連するブログ記事のクラスタから抽出される属性も因子として利用可能ではないかと考えられる。

実際にブログ記事に含まれる画像を調べてみると、意見記事を書いているケースでは消費者自身が撮影した商品画像を掲載しているケースが多く、また、商品レビューのような非意見記事ではメーカーが準備した商品写真などをそのまま（あるいはリンクして）利用しているケースが多い。この事から、商用画像をそのまま使っているのか、あるいは個人が撮影した画像を使っているのか、という特徴も意見記事と非意見記事を見分ける際のヒントにすることができる。

この際、商用画像か否かをどのようにして判別するかが問題となる。我々は，商用画像は類似するものが多く、かつしばしば同一の画像が複数の記事で利用されることに着目した。図 1 はキーワード「電子ケトル T-fal」で画像検索をした結果を類似度に基づいてクラスタリングした結果の一部である。(a), (b), (c), (d)はそれぞれ 6 枚、12 枚、1 枚、1 枚からなるクラスタである。なおクラスタリングはクラスタ数を制限するのではなく、クラスタの最大直径を制限する手法を採用しているため、一定類似度以下の画像群がクラスタを構成している。この例では(a)と(b)が商品画像から構成されるクラスタとなっており、(c)と(d)が消費者自身が購入した商品を撮影した画像となっている。このように検索結果をクラスタリングした結果、各画像が属するクラスタの大きさからその画像が商用画像かどうかある程度推測することができる。

そこで本研究では、記事中に含まれる単語、品詞の割合、アフィリエイトの有無などの情報に加え、含まれる画像が検索結果中にどの程度類似する画像を持つかという属性も用いて、意見記事と非意見記事の判別を行う手法を提案する。判定器としてサポートベクトルマシン (SVM)を用いる。本稿では上記のアイデアを裏付けるための予備実験および実験の結果について述べ、併せて現在作成中のシステムについて概要を説明する。

以下の本報告の構成は次の通りである。第 2 節では、クラスタリングで利用される画像特徴および非類似度計算方法について説明する。第 3 節では、上述の傾向を検証するための予備実験について述べる。第 4 節では、特徴語辞書と画像特徴の双方を用いて SVM による意見記事・非意見記事の分類に関する実験結果について報告する。最後に、まとめと今後の課題について述べる。

2. 画像クラスタリング

画像クラスタリングの結果を判別因子として利用するので，本節では用いているクラスタリング手法について簡単に説明を行う。画像間の非類似度を文献[8]の手法（後述）により求めて、クラスタリングを行う。クラスタリングアルゴリズムとして k - 平均法[9]を適用しているが、最適な k を求めるために、クラスタ直径が決められた範囲内に収まるように k の値を動的に変更する手法を用いている。

次に画像間の非類似度を求める手法について説明する。まず画像のカラーを HSV に変換し、HSV の各

(a)

(b)

(c)

(d)

図 1：キーワード「電気ケトル　T-fal」の検索結果をクラスタリングしたもの

チャンネル毎に平均、分散、歪みを計算する：

$$E_i = \frac{1}{N}\sum_{j=1}^{N} p_{ij}$$

$$\sigma_i = \left(\frac{1}{N}\sum_{j=1}^{N}\left(p_{ij} - E_i\right)^2\right)^{\frac{1}{2}}$$

$$s_i = \left(\frac{1}{N}\sum_{j=1}^{N}\left(p_{ij} - E_i\right)^3\right)^{\frac{1}{3}}$$

ここで、Nはピクセル数、p_{ij}はピクセルjのiチャネルの値、E_i, σ_i, s_iはそれぞれチャネルiの平均、分散、歪みである。そしてチャネルごとに得られた 3 つの値から 9 次元ベクトルを構成し、これを画像間の特徴量とする。二つの画像H,Iの平均、分散、歪みをそれぞれE_i, $\sigma_i, s_i, F_i, \zeta_i, t_i$,（i=H.S.V）とすると、これらの画像間の非類似度は次式で定義される：

$$d(H,I) = \sum_{i=H.S.V} w_{i1}\left|E_i - F_i\right| + w_{i2}\left|\sigma_i - \zeta_i\right| + w_{i3}\left|s_i - t_i\right|$$

ここで、w_{ij}（i=H,S,V;j=1,2,3）は重みである。

$$W_1 = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 1 & 1 & 1 \end{pmatrix} \quad W_2 = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 3 & 3 \end{pmatrix}$$

$$W_3 = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 4 & 4 \\ 1 & 2 & 2 \end{pmatrix}$$

文献[7]では、3 種類の重みが試されているが、本システムではW_3を用いている。これは彩度に重みを置くほうが、より、印象の似た画像により高い類似度を示す傾向が見られたためである。

3. ブログ画像と意見記事・非意見記事との関係

ブログ画像とテキストデータの両方を用いた意見文抽出手法について実験を行う前に、ブログ画像のみを用いた実験を行い、類似度に基づくクラスタリングを行った結果から抽出される属性と、意見記事・非意見記事の分類結果との間にどのような関係が成立するか、また、意見記事抽出の因子として使えるかどうかの予備実験を行った[10]。

実験は、商品を検索し手作業でレビュー記事・非レビュー記事の分類は、リンク先の記事においてブロガ自身がその商品に対して良し悪しを述べているかどうかを基準とし、手作業で分類を行った。検索キーワードとなる商品は、Amazon.co.jp と価格.com 内で人気のある商品をいくつかのカテゴリの中から選び、検索を

行った。

本実験では、商品名をキーワードとした検索結果の上位 60 件について画像が個人撮影されたものか､商用のものかという観点と、ブログ記事が意見文を含むかどうかという観点から手作業で分類した。分類結果を表 1 に示す。

表 1：分類結果(単位は%)

商品名	個人		商用	
	意見	非意見	意見	非意見
電気ケトル　T-fal	3	2	8	87
加湿器	18	8	10	63
電子辞書　Papyrus	3	0	0	97
電子辞書 EX-word	2	2	2	95
iPod nano	13	8	5	73
iPod classic	3	13	15	68
iPod touch	27	12	10	52
IXY DIGITAL 910 IS	20	27	0	53
IXY DIGITAL	3	3	10	83
LUMIX DMC-FX33	10	5	20	65
REGZA	7	8	8	77
BRAVIA	13	17	10	60
プレイステーション 3	10	8	0	82
Wii	32	33	5	30
905i	32	43	18	7
X01T	20	45	2	33
INFOBAR 2	12	35	13	40

表 1 より一部を除き、商用画像の非意見記事は検索結果に表示される画像の中に半分以上含まれることが分かる。

次に個人で撮影した画像を含みかつ意見記事であるものの個数を A、個人で撮影した画像を含みかつ非意見記事であるものの数を B、商用の画像を含みかつ意見記事であるものの数を C、商用の画像を含みかつ非意見記事であるものの数を D とする。これにより、各商品に対して個人撮影を含むブログ記事のうち意見記事を含む記事がどれだけあるか(TP)、商用画像を含むブログ記事のうち意見記事を含まないものがどれだけあるか(TN)、意見記事を含む記事のうち個人撮影画像を含むものがどれだけあるか再現率を以下の式より計算する。

$$TP = \frac{A}{A+B} \quad TN = \frac{D}{C+D} \quad \text{再現率} = \frac{A}{A+C}$$

各商品の TP、TN、再現率の計算結果を表 2 に、TP、TN、再現率の平均値を表 3 に示す。

表 2：画像の分類

商品名	TP	TN	再現率
電気ケトル　T-fal	0.67	0.91	0.29
加湿器	0.69	0.86	0.65
電子辞書　Papyrus	1.00	1.00	1.00
電子辞書　EX-word	0.50	0.98	0.50
iPod nano	0.62	0.94	0.73
iPod classic	0.20	0.82	0.18
iPod touch	0.70	0.84	0.73
IXY DIGITAL 910 IS	0.43	1.00	1.00
IXY DIGITAL	0.50	0.89	0.25
LUMIX DMC-FX33	0.67	0.76	0.33
REGZA	0.44	0.90	0.44
BRAVIA	0.44	0.86	0.57
プレイステーション 3	0.55	1.00	1.00
Wii	0.49	0.86	0.86
905i	0.42	0.27	0.63
X01T	0.31	0.95	0.92
INFOBAR 2	0.25	0.75	0.47

表 3：評価尺度の平均

	TP	TN	再現率
平均	0.52	0.86	0.62

表 3 より TN に関しては平均が 8 割を超える値となり、商用画像を用いたブログページが非意記事である可能性が高いことが判明した。しかし、再現率に関しては、6 割程度の結果となり、個人撮影の画像が含まれるブログページを取り出してしまうと意見記事を見落としてしまう可能性が高いということが分かる。

ただし、表 2 から見て取れる通り商品ごとに有効性についてはばらつきが著しい。したがって、平均値は必ずしも良いとは言えないが、分類パラメータの一つとして利用した時に有効に機能するケースもあると考えられる。

表 1～3 より、商用画像による非意見文の除去は TN の平均値が 8 割以上であることから、良い結果が得られたと考えられる。しかし、意見文抽出に関しては TP や再現率が 5～6 割という結果から､あまり精度は高くないと言える。また､「Wii」や「905i」といった趣味の要素が大きい商品に関する検索結果は撮影した画像が多い結果となり、画像のみでは意見文抽出は難しいと考えられるが、ブログ画像と SVM によるテキスト

データを用いた意見記事抽出において、精度が向上する可能性は充分あると考えられる。

4. 意見記事判別実験

ブログ画像とテキストデータの両方を用いた意見記事抽出手法について実験を行った。記事の分類にはTinySVM[11]を用い、訓練データとしては画像が含まれるHDDレコーダー（DIGA, VARDIA, 楽レコ, スグレコ）、電気ケトル（T-fal, ラッセルホブス, タイガー）、デジタルカメラ（Cyber-shot, EXILIM, FinePix, IXY DIGITAL, LUMIX, Optio）、携帯音楽プレイヤー（iPod, WALKMAN, ZEN）、サプリメント（アミノバイタル, ウィダー, ネイチャーメイド）、TV（AQUOS,BRAVIA, REAL, REGZA,VIERA, Wooo）に関する記事をそれぞれ100件（意見記事50件、非意見記事50件）収集して利用した。検索エンジン（Yahoo!ブログ検索[12]、Yahoo!画像検索[13]）で提供される画像検索機能を利用した。

SVMに用いる属性としては文献[7]を参考にし、基本属性としては特徴語辞書、品詞の割合、アフィリエイトリンクの有無を、また、追加の属性として非類似度に基づくクラスタリングを行った結果から抽出される属性を用いた。特徴語辞書はAmazon Webサービス[14]を用い、通販サイトAmazon.co.jpからＨＤＤレコーダー、電気ケトル、デジタルカメラ、携帯音楽プレイヤー、サプリメント、ＴＶの商品説明とカスタマレビューに含まれる動詞、感動詞、形容詞、形容動詞を取得し、商品説明のみに含まれる単語と、カスタマレビューにのみ含まれる単語をドメインごとにまとめ特徴語辞書とした。単語の切り出しには茶筌[15]を用いた。

特徴語としては、例えばデジタルカメラの場合、カスタマレビューからは、「凹む」、「かさ張る」、「あきれる」、「風変わり」など773単語、商品説明からは、「カジュアル」、「高密度」、「ドラマチック」、「鮮烈」など331単語が抽出された。また、品詞の割合としては名詞、動詞、形容詞の出現頻度をそれぞれN、V、Aとした時、N/(N+V+A)、V/(N+V+A)、A/(N+V+A)とした。

画像特徴はクラスタリング結果から算出する。算出方法は次の通りである。まず、訓練データに含まれる画像群を、非類似度を元にクラスタリングする。次に画像毎にそれが属するクラスタが複数の画像を含んでいるか否かを求める。複数ある場合にはそうでない場合には0を付与した。なお、実際に利用する際には画像検索結果に含まれる画像群がクラスタリングの対象となる。

本実験では、各ドメインの収集データに対してleave-one-outによる交差検定を行い、正確性、適合率、再現率を求めた(以下の式を参照)。なお、テスト対象となるケースの画像特徴については収集データ全体に対して上述の処理を行って得られた値を利用した。結果を表4に示す。

$$\text{正確性}=\frac{A+D}{A+B+C+D} \qquad \text{適合率}=\frac{A}{A+C} \qquad \text{再現率}=\frac{A}{A+B}$$

表4：記事の分類結果

ドメイン	画像特徴使用の有無	正確性（%）	適合率（%）	再現率（%）
HDDレコーダー	有り	82.0	88.1	74.0
	無し	71.0	88.9	48.0
電気ケトル	有り	71.0	83.9	52.0
	無し	53.0	80.0	8.0
デジタルカメラ	有り	61.0	86.7	26.0
	無し	72.0	80.6	58.0
携帯音楽プレイヤー	有り	66.0	66.7	64.0
	無し	67.0	66.7	68.0
サプリメント	有り	69.0	67.9	72.0
	無し	70.0	67.9	76.0
TV	有り	48.0	47.8	44.0
	無し	58.0	56.7	68.0

表4 HDDレコーダー、電気ケトル、デジタルカメラでは正確性、適合率、再現率のいずれかで性能の向上が見られた。HDDレコーダーと電気ケトルでは正確性、再現率の大幅な向上が見られた。デジタルカメラでは適合率が上がる代わりに正確性と再現率が低下している。これらに対し、性能の低下のみが見られたのが携帯音楽プレイヤー、サプリメント、TVである。携帯音楽プレイヤーとサプリメントでは正確性と再現率が若干ではあるが低下してしまっている。またTVでは大幅な性能の低下が見られる。検証実験での結果が示すとおり、今回採用した画像特徴はドメインごとの有効性にばらつきがあるためこのような結果になったと考えられる。

以上より、画像特徴を用いることでいずれかの指標を向上させることが出来たドメインが5割あることがわかった。これに既存手法と同程度の結果が出ているドメインを加えると、実験を行った約8割のドメインに達した。このことから画像をクラスタリングした結果から抽出される属性を加えることで有効に機能するドメインがあることを示した。

しかし、ドメインごとにばらつきがあるため実際にシステムとして実装する際には意見記事として判別されたものだけでなく、非意見記事として判別されたも

のも利用者に提示する必要があると考えられる。

5. 意見記事判定システムの設計と実装

意見記事・非意見記事判定システムは利用者と検索エンジンの橋渡しを行うメディエータである。システムでは、利用者が入力した検索単語を元に記事をおよび記事に含まれる画像を取得し、取得した記事と画像から SVM のテストデータを作成する。作成したテストデータを SVM にかけ、意見記事・非意見記事を判別した結果を利用者に返す。なお、実装したシステムでは SVM の判別に用いる属性として、特徴後、品詞の割合、画像特徴を用いている。アフィリエイトリンクに関する属性を利用することも考えられるが、本システムでは実装にいたっていない。システムのシーケンス図を図 2 に示す。

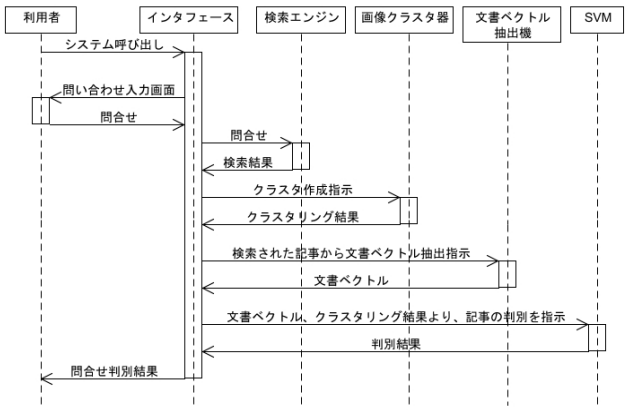

図 2：システムのシーケンス図

検索エンジンとしては Yahoo!画像検索を用いた。Yahoo!ブログ検索を用いることも考えられるが、検索キーワードと画像の関連度を判定する処理が複雑になるため本システムでは Yahoo!画像検索を用いている。したがって、判別の対象には一般の Web 日記なども含まれる。あえてブログ記事のみに限定したい場合は「Blog」などのキーワードを追加することである程度可能である。なお、本システムでは記事本文を抽出する際に RSS フィードを利用しているため、メタデータを持たないページは意見記事判定の対象からは外れる。

6. 利用例

実装したシステムの入力画面を図 3 に示す。

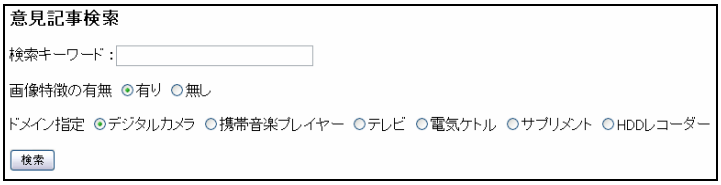

図 3：システムの入力画面

入力画面では検索に用いるキーワード、画像特徴の有無と判別に使用するドメインを利用者が指定する。また、表 4 よりドメインによっては画像特徴を用いた場合の方が判別の性能が悪くなってしまうことがわかっている。そこで、システムにおいて画像特徴を使用するか否かは利用者自身によって選択できるようになっている。利用できるドメインは訓練データとして集めた HDD レコーダー、電気ケトル、デジタルカメラ、携帯音楽プレイヤー、サプリメント、TV の 6 つである。検索対象に合ったドメインを指定することでより適した学習データを用いることができる。

図 3 の画面に対して、利用者が検索キーワードとして「レコーダー　VARDIA」と入力し、画像特徴の有無を「有り」、ドメインとして「HDD レコーダー」を選択して検索ボタンをクリックしたとする。検索結果の記事を判別した結果が利用者に返される。執筆時点では、意見記事が 4 件、非意見記事が 9 件、判定不能記事が 87 件である。適合率は 100%で再現率は 66.7%であった。判別結果のうち意見記事の提示部を図 4 に、非意見記事の提示部を図 5、6 に、判定不能記事の提示部の一部を図 7 に示す。

図 4：意見記事提示部

図 5：非意見記事提示部(パート 1)

図 6：非意見記事提示部(パート 2)

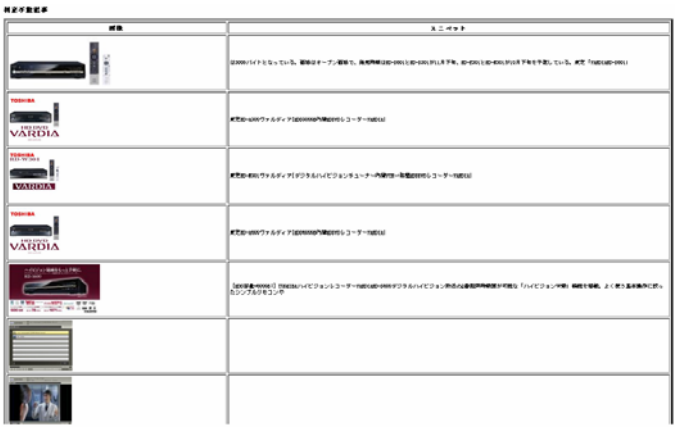

図 7：判定不能記事の提示部（一部）

7. まとめと今後の課題

本研究では、ブログに含まれる画像データから抽出される属性と既存研究で用いられている属性(特徴語、品詞の割合、アフィリエイトリンク)を組み合わせたハイブリッドな意見記事判定処理手法の提案をした。

ブログ記事に含まれる画像を調べてみると、意見記事を書いているケースでは消費者自身が撮影した商品画像を掲載しているケースが多く、また、商品レビューのような非意見記事ではメーカーが準備した商品写真などをそのまま（あるいはリンクして）利用しているケースが多い。この事から、商用画像をそのまま使っているのか、あるいは個人が撮影した画像を使っているのか、という特徴も意見記事と非意見記事を見分ける際のヒントにすることができると考えた。

そこで、ブログ画像をクラスタリングした結果と、意見記事・非意見記事との間にどのような関係が成立するか、また、意見記事抽出の因子として使えるかどうか検証するために実験を行った。検証実験の結果、表 3 より TN に関しては平均が 8 割を超える値となり、商用画像を用いたブログページが非意記事である可能性が高いことが判明した。しかし、再現率に関しては、6 割程度の結果となり、個人撮影の画像が含まれるブログページを取り出してしまうと意見記事を見落としてしまう可能性が高いということが分かる。

ただし、表 2 から見て取れる通り商品ごとの有効性についてはばらつきが著しい。したがって、平均値は必ずしも良いとは言えないが、分類パラメータの一つとして利用した時に有効に機能するケースもあると考えられる。

検証実験の結果を基にブログ画像とブログのテキストデータの両方を用いた意見記事・非意見記事の判別実験を行った。実験の結果から、画像特徴を用いることでいずれかの指標を向上させることが出来たドメインが 5 割あることがわかった。これに既存手法と同程度の結果が出ているドメインを加えると、実験を行った約 8 割のドメインに達した。このことから画像をクラスタリングした結果から抽出される属性を加えることで有効に機能するドメインがあることがわかった。

最後にブログ画像とブログのテキストデータの両方を用いた意見記事・非意見記事判定システムを web アプリケーションとして実装した。実装したシステムを用いることで利用者が興味のある商品のユーザレビューを簡単に収集することが出来る。

今後の課題は次の通りである。まず、特徴語辞書による判定の精度そのものが必ずしも良好ではないため、辞書の作成方法について再検討が必要である。なお、ブログ記事は記事中の単語数が極端に少ないことがあり、これが辞書に基づく判別を難しくしている点は他の研究と同様である。次に、本研究では SVM を用いたのであるが、そのパラメータについてはほぼデフォルトのものを用いておりチューニングがなされていない。他の判別器の利用も含め、改めて検討することが必要である。

また、本稿での実験では訓練データと特徴語辞書を

同一ドメインとしているが、訓練データと特徴語辞書が異なるドメインの実験を行う必要がある。異なるドメインを用いた場合でも良好な結果が出れば、実装したシステムにおいて大域的なドメイン選択が可能になる。現在は実験に用いたドメインである HDD レコーダー、電気ケトル、デジタルカメラ、携帯音楽プレイヤー、サプリメント、TV の中から利用者が選択するようになっているが、かなり小さなドメインであるため利用しづらくなっている。このことを改善するためにも、訓練データと特徴語辞書が異なるドメインの実験を行い、大域的なドメインを作成する必要がある。

謝辞

本研究を行うにあたり、有益なご助言をいただいたソフトウェアシステム工学研究室の方々に深く感謝します。

参 考 文 献

[1] L. Page, S. Brin, R. Motwani and T. Winograd, “The PageRank citation ranking: Bringing order to the Web,” Stanford Digital Library Technologies Project Working Paper, 1997-0072.

[2] J. Kleinberg, “Authoritative sources in a hyperlinked environment,” Technical Report RJ 10076, TBM, May 1997.

[3] M. E. Newman, “Fast algorithm for detecting community structure in networks,” Physical Review E, 60, 2004, p.066133.

[4] 奥村学, “ブログマイニング,” 情報処理学会研究報告, 2006-DBS-139(5), 2006, pp.33-44.

[5] 松永拓, 平手勇宇, 山名早人, “キーワードの出現に基づくブログコミュニティ抽出とオピニオンリーダーの発見,” 第 18 回データ工学ワークショップ（DEWS2007）, 広島, Feb.28-Mar.2, 2007.

[6] 高木允, 森康真, 田村慶一, 黒木進, 北上始, “ブログユーザ空間からの頻出なコミュニティ抽出法,” 第 18 回データ工学ワークショップ（DEWS2007）, 広島, Feb.28-Mar.2, 2007.

[7] 川口敏広, 松井藤五郎, 大和田勇人. “SVM と新聞記事を用いた Weblog からの意見文抽出,” 人工知能学会第 20 回全国大会(JSAI2006), 東京, June.7-9, 1A3-3(2006).

[8] M. A. Stricker and M. Orengo, “Similarity of color images,” Storage and Retrieval for Image and Video Database(SPIE), pp.381-392(1995).

[9] R. O. Duda, P. E. Hart and D. G. Stork, “Pattern Classification/2ed,” Wiley(2001).

[10] 桑原卓朗, “blog 画像情報を用いた意見・評判情報抽出手法の提案,” 山口大学卒業論文, 2008.

[11] 奈良先端科学技術大学院大学 松本研究室, “TinySVM: Support Vector Machines” http://chasen.org/~taku/software/TinySVM/.

[12] Yahoo!Japan, “Yahoo! ブログ検索,” http://blog-search.yahoo.co.jp/.

[13] Yahoo!Japan, “Yahoo! 画像検索,” http://image-search.yahoo.co.jp/.

[14] Amazon.com, Inc., “Amazon Web サービス,” http://www.amazon.co.jp/gp/feature.html?docId=451209.

[15] 奈良先端科学技術大学院大学, “形態素解析システム茶筌,” http://chasen.naist.jp/hiki/ChaSen/.

付録

A. 利用した画像の URL

図 1 で利用している画像を掲載しているページ URL は本論文執筆時において次の通りである。

(a)http://www.goods5.jp/goods/tfal/

http://store.shopping.yahoo.co.jp/click/seb-015.html

http://www.citibank.co.jp/ccsi/direct/point/cpoint/kitchen.html

http://depart.livedoor.com/

http://www.t-fal.co.jp/tefal/services/call.asp

http://ata-xyz.blog.so-net.ne.jp/2007-06-02

(b)http://www.t-fal.co.jp/tefal/products/family/410/justin.asp

http://www.888123.co.jp/kis/044/0370.htm

http://www.murauchi.com/MCJ-front-web/CoD/0000000539209

http://www.seikatu-sutairu.com/2006/11/_12l.html

http://www.888123.co.jp/SHOP/044-1019.html

http://pecopeco.seesaa.net/archives/200507-1.html

http://plaza.rakuten.co.jp/kafunkaden/diary/20061002

http://www.tim.hi-ho.ne.jp/%20wec7/mono_benrikaden.htm

http://citygas.co.jp/ataru/

http://www.murauchi.com/MCJ-front-web/CoD/0000000834149

http://store.shopping.yahoo.co.jp/click/seb-027.html

http://shop.prokitchen.co.jp/item_list.command?category_cd=KK_DENKIPOD

(c)http://bits-blog.seesaa.net/article/1908207.html

(d) http://avant.mo-blog.jp/private_room/2006/01/tf_a271.html