買い物Blogに対する機械学習を用いた自動評価分析法の提案

2003MT121 加藤 圭喬
指導教員　河野 浩之

1　はじめに

近年, Weblog(Blog) コンテンツが急増している．Blogの量が増えるにつれ，興味のある Blog を検索・収集することが難しくなってきた．そこで，検索を少しでも容易にするための手法として，Blog 記述の内容に従って，特定のカテゴリに分類するという方法がある．これを行うため，テキストマイニングを用いた研究が行われている [1]．この研究では Blog 記述などの本文のみを参照することで分類をしているが，他の属性を考慮した分類は行われていない．そこで，本研究では位置情報・記述者情報を属性として考慮して分類を行う．この操作を行うことで，分類精度の向上を目指す．

自動評価分析では，分類結果から Blog で記述された対象となる地点の位置情報（緯度・経度）および記述者 ID を取得し，この情報を元に各地点における評価および Blog 記述者ごとによる評価のパターンを Google Maps を用いて視覚化する．

2　評価表現や単語の出現頻度を用いた感情分類法

評価分類を行う関連研究として，感情分類法という研究が行われている．これは，記述内容がポジティブか，ネガティブか，中立表現なのか自動的に分類する手法である．

Suzuki ら [2] は Semi-supervised 学習法による評価表現抽出と分類を行っている．評価表現とは，記述における意見・主張の主要部分を「対象／属性／評価語」の 3 要素の組で構成されると定義している．つまり，どれ＜対象＞についての，どの部分＜属性＞が，どうなのか＜評価語＞を抽出している．こうして，辞書を作成し，評価分類を行っている．さらに，評価表現がある特定の周辺情報（記述）を伴って出現すると仮定することで，評価表現・周辺情報を得るたびに，未知の新しい評価表現をより正しく分類できるようにナイーブベイズ分類器・SVM・EM アルゴリズムを組合せ，実験している．

藤村ら [3] は C4.5 による決定木による分類学習アルゴリズムを用いた実験を行っている．素性には単語の出現頻度を用いおり，この出現頻度の問題を扱うため，TF*IDF 法を用いている．彼らの提案するアルゴリズムはノート PC に関する掲示板 (http://kakaku.com/) の 2850 個のコメントからなるコーパスを対象に適用され，良・悪の評判に分類している．

3　自動評価分析法の提案

3.1　データの準備と前処理

買い物に関する Blog データを用いる．自動評価分類を行うためには，記述者 ID，買い物場所の位置情報，タイトル，本文のデータをもった Blog を用意する必要がある．用意した Blog データから評価語を抽出する．そして，ポジティブ／ネガティブ／中立のいずれかのラベルをつけ，評価語辞書を作る．

続いて，Blog 記述の形態素解析を行う．茶筅 (ver.2.3.3)[*1]を利用する．茶筅は奈良先端科学技術大学大学院で開発されたプログラムである．解析された記述は，それぞれ，見出し語ごとに分かち書きする．例えば，「このケーキの生クリームは美味しい.」という記述が与えられた場合，以下のように分かち書きがなされる．
「この｜ケーキ｜の｜生クリーム｜は｜美味しい｜.」
「連体詞｜名詞｜助詞｜名詞｜助詞｜形容詞｜ 記号」

さらに，Blog 記述ごとに，文書ベクトルを与える．文書ベクトルは，分かち書きされた記述と評価語辞書に登録された評価語とを照らし合わせて作成する．ベクトルはそれぞれポジティブ，中立，ネガティブと 3 つの成分を持たせる．Blog 記述内で各評価語が出るたびに，それぞれの成分に値 “ w_i ” を与える．この値は先行研究 [3] と同様に，TF*IDF 法によって重みをつけて与える．上記の例で考えると，「美味しい」という評価語が辞書に登録されていた場合，文書ベクトルのポジティブの要素に値が加えられる．このベクトル用いて分類をする．

3.2　機械学習による分類の評価

データマイニングツール WEKA(ver.3.4.8a)[*2]を用いる．WEKA は，Waikato 大学が中心となって開発しているツールである．

分類学習アルゴリズムを作るため，先行研究 [2,3] で使用された C4.5 による決定木，ナイーブベイズ分類器 (NB)，Support Vector Machine(SVM) を用いて学習モデルを作成する．この学習モデルを用いて，Blog 記事全体の分類評価を行う．

3.3　評価分析の手順

分類された Blog データに対し，各地点における評価のパターンを分析する．ここで位置情報とは，施設コー

[*1] Chasen’s Wiki, http://chasen.naist.jp/hiki/ChaSen/
[*2] WEKA, http://www.cs.waikato.ac.nz/ml/weka/

ドおよび緯度・経度によって示される．各 Blog 記述で評価された施設・広告看板等の位置情報を読み込み，同一地点における評価の数をスコアリングする．緯度経度，施設名，フロア，各評価の数，記述した性別の数，最も評価の多かった分類のラベル (ポジティブ：+1，中立：0，ネガティブ：- 1)，最も多かった性別のラベル（男性：+1，中性：0，女性：-1）を CSV ファイルに書きこむ．この操作を全ユーザとユーザ一人一人を対象に行う．

3.4 評価分析の視覚化

各座標におけるスコアを地図上で可視化する．ここで，本研究では Google Maps を用いて視覚化処理を行う．Google Maps は Google.com から提供される地図情報サービスである．Web ページ上に組み込み，JavaScript により加工することができる．視覚化システムの機能には次の 3 つの機能と手順を用意する．

1. 全記述者による，指定された施設における評価と Blog 記事の表示
2. ユーザー個人による，全施設の評価と Blog 記事の表示
3. 指定された施設における性別的な評価傾向と Blog 記事の表示

4 分類・分析実験と評価

本研究では，東京大学の羽藤助教授の収集された 4985 個の Blog データを用意した．評価するにあたり，あらかじめ記述ごとに評価のラベルを与えた．また，このデータから評価語を抽出し，辞書を作成した．続いて，文書ベクトルを与えるため，データを文書ベクトル作成プログラムにかける．このうち 100 個の Blog 記事を訓練データとし，分類アルゴリズムを用いて学習させる．学習時に使用する属性値は文書ベクトルの各要素で分類先は評価ラベルである．この学習モデルを Blog 記事全てに対し適用し，分類精度を測った．この結果，各分類アルゴリズムによる分類精度は，表 1 のようになった．

表 1　評価語のラベル付けの割合

アルゴリズム	分類精度
C4.5 決定木	84.49%
NB	80.84%
SVM	85.07%

また，施設コード，ユーザ ID ごとにデータを分割し，それぞれ個別に分類精度を測った．この結果，施設コードでは駅のような慌しい施設での記述は直感的な記述になることが多く，分類精度が高くなった．また，ユーザ ID では食品に対する評価を率直に行っていた記述者は，高精度となる傾向が見られた．

このように，記述場所，記述者ごとに特性が見られ，90% 以上の精度のものも多く，位置情報・記述者情報を属性として考慮することで精度の向上が見られた．

そして，評価分析システムを実装し実行したところ，図 1 のような出力が得られた．

図 1　評価分析視覚化システム

5 まとめ

本研究では，買い物に関する Blog を用意し，その内容を分析して位置情報・記述者情報を属性として考慮した分類評価を行った．この情報を考慮することで，90% 以上の精度となったものがあり，精度の向上が見られた．

また評価分析の視覚化システムの実装も行った．

参考文献

[1] 鈴木泰裕, 高村大也, 奥村学, “ Weblog を対象とした評価表現抽出, ” 第 6 回セマンティックウェブとオントロジー研究会 (SIG-SW&ONT-A401-02), pp.1-10, 2004.

[2] Y. Suzuki, H. Takamura, M. Okumura, “ Application of Semi-supervised Learning to Evaluative Expression Classification, ” Proc. of the 7th International Conference on Intelligent Text Processing and Computational Linguistics CICLing-2006, pp.502-513, 2006.

[3] 藤村滋, 松村真宏, 岡崎直観, 石塚満, “ 電子掲示板上の評判情報に基づく意思決定支援, ” 第 17 回 人工知能学会全国大会 (JSAI2003), 2B1-05, pp.1-2, 2003.