意味を考慮したBlog記事の分類に関する研究

中村健二† 田中成典‡ 細畠啓史† 吉村智史† 北野光一† 大谷和史‡
関西大学大学院総合情報学研究科† 関西大学総合情報学部‡

1．はじめに

近年，Web上では，Blogの開設数が増加[1]している．Blogには，消費者の商品に関する意見が掲載されることもあるため，企業は，Blogから消費者の意見を収集することで，市場の調査や商品の改善に役立てることができる．しかし，近年のBlog数の急激な増加により，Blog記事の中から必要な情報のみを効率よく取得することは困難である．そのため，既存研究では，Blog記事群等の文書集合を文書中の単語に基づいて分類する研究[2][3]やBlog記事から記事内容を代表する単語を抽出する研究[4][5]が行われている．しかし，これらの研究では，Blog記事が所属する分野による単語の意味の違いを考慮しておらず，分野ごとの正確な分類結果を得ることができないという問題がある．そこで，本研究では，Blog記事から抽出した単語とBlog記事が所属する分野を用いて，Blog記事の分野による単語の意味の違いを考慮した分類を目指す．

2．システムの概要

本研究では，トピックの分野を考慮したBlog記事群の分類手法を提案する．システムの概要を図1に示す．本システムは，1）データベース構築機能，2）トピックベクトル作成機能，3）トピック分類機能により構成される．入力データは，分析対象商品名とし，出力データは，整理したBlog記事群とする．

2．1　データベース構築機能

本機能では，Blog記事収集処理と概念ベース構築処理の2つの処理を行う．Blog記事収集処理では，Web上に存在するBlog記事を自動的に取得する．概念ベース構築処理では，まず，Blog記事データベースからトピックの分野に所属するBlog記事を取得する．概念ベース構築処理では，トピックの分類に影響を与えない素性

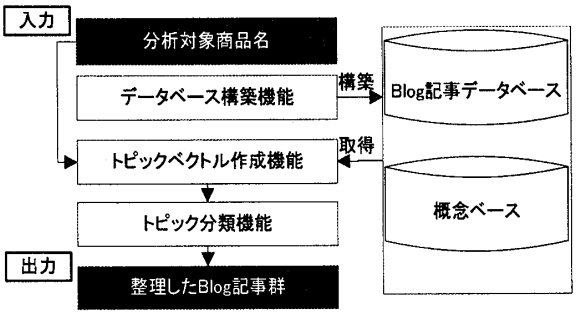

図1　システムの流れ

を省くため，記事の形態素についてトピックの分野を特徴付ける値である情報利得値を算出し，情報利得値が高い形態素の出現頻度を用いて形態素の重みを算出する．

2．2　トピックベクトル作成機能

本機能では，入力した分析対象商品情報に関するBlog記事群からトピックを表すトピックベクトルを作成する．まず，Blog記事データベースから入力した分析対象商品情報に関するBlog記事群を取得する．次に，そのBlog記事のトピックから所属する分野を判定し，参照すべき概念ベースを決定する．最後に，Blog記事における単語の出現頻度から各トピックのトピックベクトルを作成し，作成したトピックが参照する概念ベースを用いて算出した値でトピックベクトルの重みを再計算する．

2．3　トピック分類機能

本機能では，トピックベクトルを分類する．まず，トピックベクトルとクラスタの類似度をベクトル空間モデルのコサイン尺度で算出し，類似度が閾値以上ならばトピックベクトルをクラスタに分類する．ここで，クラスタが存在しない場合と類似度が閾値以上のクラスタが存在しない場合は，クラスタを新たに作成する．また，クラスタとトピックベクトルの類似度の算出には，クラスタのベクトルを決定するために最長距離法を用いる．次に，この処理を全クラスタに対して繰り返すことでトピックベクトルを分類する．最後に，各クラスタに含まれる記事を並べて整理したBlog記事群を表示する．

Research for Classifying Blog Entries with Semantic Information
†Kenji Nakamura, Hirofumi Hosohata, Satoshi Yoshimura, Koichi Kitano
Graduate School of Informatics, Kansai University, 2-1-1 Ryouzenji-cho Takatsuki-shi, Osaka 569-1095, Japan
‡Shigenori Tanaka, Kazufumi Ohtani
Faculty of Informatics, Kansai University, 2-1-1 Ryouzenji-cho Takatsuki-shi, Osaka 569-1095, Japan

3．システムの実証実験と考察

本システムの実行結果を図2に示す．概念ベースを使用して分類精度を向上させる本提案手法の有用性を検証するため，本システムにおいて概念ベースを使用した場合と概念ベースを使用しない場合で分類精度の比較実験を行った．

3．1　実証実験

本実験で用いる Blog 記事群は，Yahoo!ブログ内の 2007/12/01 から 2007/12/20 の間に書かれた Blog 記事とし，対象とするトピックの分野は Yahoo!ブログが提供する分野を用いた．トピックベクトルの分類に用いるコサイン尺度による類似度の閾値は，実験を繰り返し，偏ったグループが生成されないように最適な値として 0.02 を採用した．本実験では，評価指標として各グループにおいて F 値を算出した．F 値とは，適合率と再現率の調和平均であり，分類精度の総合的な指標として用いられる．適合率とは，実際に分類された情報のうち正しい情報の割合である．また，再現率とは，文書集合全体の正しい情報のうち実際に分類された情報の割合である．本実験では，分析対象の商品によって結果が異なる可能性があるため，複数の商品に対して実験を行った結果の平均値を採用した．

3．2　結果と考察

本システムにおける概念ベースを使用した場合と概念ベースを使用しない場合を比較することで表1に示す結果が得られた．この結果から，概念ベースを使用した場合は概念ベースを使用しない場合に比べて F 値が高いことがわかる．適合率については，本手法の方が高い値が得られた．これは，概念ベースを用いることにより，トピックベクトルが Blog 記事を特徴付けることができるようになったためであると考えられる．再現率については，既存手法の方が高い結果となった．これは，概念ベース内に存在しない語を多く含む Blog 記事において，相対的に特徴が薄れたことにより，どのクラスタにも属さないトピックベクトルの出現頻度が増加したことが原因であると考えられる．この実験結果により，文書中の単語情報のみを手がかりとして分類するだけではなく，各トピック分野における概念ベースを用いることにより，単語の意味の違いを考慮して Blog 記事を分類する本提案手法が有用であることが分かった．

4．おわりに

本研究では，意味を考慮した Blog 記事群の分類手法を考案した．実証実験の結果，Blog 記事内の出現単語だけではなく，Blog 記事の所属する分野を使用し，Blog 記事の意味付けの精度を

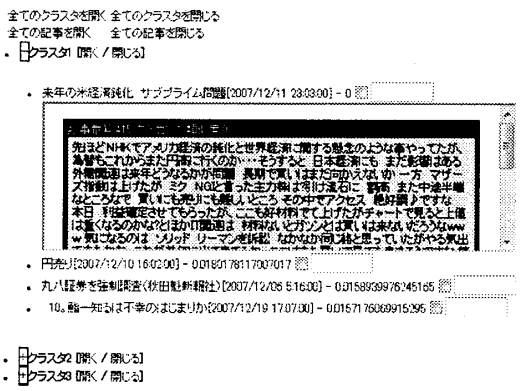

図2　本システムの実行結果

表1　実験結果

評価項目	概念ベース使用	概念ベース非使用
適合率	0.5873	0.4228
再現率	0.4532	0.4932
F値	0.5116	0.4552

向上できた．これにより，本手法の有用性を実証できた．しかし，本研究では，Blog 記事の所属する分野をあらかじめ設定する必要があるため，様々なトピックが扱われる Blog への対応が不十分である．今後は，Blog 間の関連性を考慮することにより固定された分野に捕らわれない手法の考案および、分類精度を向上させるために手法の改良を行う予定である．

参考文献

[1] 総務省：平成 18 年度版情報通信白書，ぎょうせい，2006.7.

[2] 平野耕一，古林紀哉，高橋淳一：日本語圏ブログの自動分類，自然言語処理研究会研究報告，情報処理学会，Vol.2005，No.117，pp.21-26，2005.11.

[3] Joachims，T.：Text Categorization with Support Vector Machines；Learning with Many Relevant Features，Processing of the 10th European Conference on Machine Learning，pp.137-142，1998.3.

[4] 関口裕一郎，佐藤吉秀，川島晴美，奥田英範，奥雅博：blog ページ集合に対する話題語句抽出手法，自然言語処理研究会研究報告，情報処理学会，Vol.2005，No.117，pp.27-32，2005.11.

[5] Kleinberg，J.：Bursty and Hierarchical Structure in Streams，Proceeding of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，ACM Press，pp.91-101，2002.3.