1T-5

ベイジアンフィルタとカテゴリ分類を用いたブログスパム判定に関する研究

中村健二† 田中成典‡ 池辺正典† 吉村智史† 寺口敏生‡
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，誰もが気軽に情報発信を行う手段として，ブログ（Weblog）が注目されている．ブログは，記事，コメントとトラックバックから構成され，コメントを通じた情報交換やトラックバックを通じた関連情報の収集が容易である．しかし，それらの機能を対象として，広告や他サイトへの誘導を目的とするスパム投稿が増加[1]しており，投稿のスパム判定を効果的に行う方法が求められている．既存研究として，同一文字列を検出することでスパム判定を行う研究[2]や，ベイジアンフィルタを適用し，文中で用いられる単語の頻度からスパム判定を行う研究[3]などがある．しかし，同一文字列からスパム判定を行う手法は，誤判定が多いという問題がある．また，ベイジアンフィルタによる手法は，判定対象のブログ記事の内容を考慮しないため，記事のカテゴリによっては，精度が低下するという問題がある．そこで，本研究では，カテゴリを分類して記事を管理するというブログの特性を活かし，カテゴリ分類[4]を用いたベイジアンフィルタ[5]によって，ブログに適したスパム判定を行うシステムを開発する．

２．システムの概要

本研究では，ブログデータを入力し，使用される単語の各カテゴリにおける出現頻度の違いから，コメントやトラックバック元の記事でのスパム判定の結果を出力する．本システムの流れを図１に示す．本システムは，１）ブログ解析機能，２）判定情報作成機能，３）スパム判定機能の３つの機能により構成される．

２．１ ブログ解析機能

本機能では，教師データとして収集したブログデータを HTML 解析し，ブログ記事，コメント及びトラックバック元のブログ記事に分割する．教師データは，既にカテゴリ分けされているブログを対象とした．

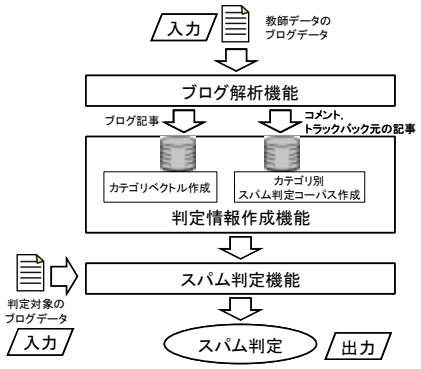

図１ システムの流れ

２．２ 判定情報作成機能

本機能では，ブログ記事のカテゴリ判定に用いるカテゴリベクトルの作成と，コメントやトラックバック元の記事のスパム判定に用いるカテゴリ別スパム判定コーパスの作成を行う．カテゴリベクトルは，既にカテゴリ分けされているブログ記事の中に出現する各単語の出現回数をもとに作成する．カテゴリベクトルの各要素の値は，出現分野の偏りを考慮して相互情報量を用いて算出する．カテゴリ別スパム判定コーパスは，カテゴリ分けされているブログ記事に投稿されたコメントやトラックバック元の記事を手動でスパム投稿とスパムではない投稿に分け，各単語の出現頻度を算出する．その結果から，スパム投稿とスパムではない投稿の比率を用いて各単語のスパム確率を算出し，スパム判定コーパスを作成する．

２．３ スパム判定機能

本機能では，まず，判定情報作成機能で作成したカテゴリベクトルとブログ記事の記事ベクトルの類似度からカテゴリ判定を行う．次に，カテゴリ判定の結果から，利用するスパム判定コーパスを決定する．最後に，カテゴリに対応したスパム判定コーパスを用いて，コメントやトラックバック元の記事のスパム確率を算出し，投稿されたコメントやトラックバック元の記事

Research on Detection Method of Blog Spam with Bayesian Filter and Categorization
†Kenji Nakamura, Masanori Ikebe, Satoshi Yoshimura
Graduate School of Informatics, Kansai University, 2-1-1 Ryouzenji-cho Takatsuki-shi, Osaka 569-1095, Japan
‡Shigenori Tanaka, Toshio Teraguchi
Faculty of Informatics, Kansai University, 2-1-1 Ryouzenji-cho Takatsuki-shi, Osaka 569-1095, Japan

のスパム判定を行う．スパム確率算出には，次に示す Gary Robinson 方式を用いた．

$$H = C^{-1}\left(-2\log\prod_{w}(1-f(w)),2n\right) \quad (1)$$

$$S = C^{-1}\left(-2\log\prod_{w}f(w),2n\right) \quad (2)$$

$$I = \frac{1+H-S}{2} \quad (3)$$

式（1）～（3）において，C^{-1} は逆カイの２乗関数を表す．また，$f(w)$は単語 w のスパム確率，n は出現単語数，Hはノンスパム性，Sはスパム性を表し，Iは Sと Hを統合した指標である．このとき，0.5 を基準として，Iが１に近ければスパムと判定し，０に近ければスパムではないと判定する．

３．システムの実証実験と考察

実証実験では，本システムの有用性を確認するため，カテゴリを考慮したスパム判定コーパスを用いる提案手法と，カテゴリを考慮しないスパム判定コーパスを用いる従来手法との比較実験を行った．

３．１　実証実験

実証実験として，本提案手法と従来手法の比較を行った．まず，カテゴリベクトル作成処理では，Yahoo!ブログの 12 カテゴリを分類対象とし，それぞれ 3,000 件の教師データを用意する．そして，この教師データをもとにカテゴリベクトルを作成する．カテゴリ判定には，ベクトル空間モデルを用いた．次に，スパム判定コーパス作成処理では，カテゴリ毎に，スパム投稿 500 件とスパムでない投稿 500 件の合計 1,000 件のコメントとトラックバック元の記事を教師データとして用意した．そして，これらのデータをもとに，提案手法のカテゴリ別スパム判定コーパスと従来手法のスパム判定コーパスを作成した．最後に，コメントやトラックバックが１件以上投稿されているブログ記事 100 件を実験対象とし，判定精度の比較実験を行った．

３．２　結果と考察

実証実験の結果を表１に示す．ここで，表１において，判定確率はスパム投稿を正しく判定する確率，擬陽性率はスパムでない投稿を誤判定する確率，擬陰性率はスパム投稿を誤判定する確率を示す．実験結果では，提案手法は従来手法に比べ，判定確率は高く，擬陽性率と擬陰性率は低くなった．スパム判定では，判定確率が高く，擬陽性率と擬陰性率が低いことが望ましいため，提案手法は従来手法よりもブログスパムの判定に適していることが実証された．
しかし，スパム判定としては，判定確率は 70%未満と低く，擬陽性率と擬陰性率は 10%以上の高い値をとっている．これは，コメントやトラックバック元の記事に出現する単語数が少なく，スパム判定コーパスの性能が十分に発揮できないためと思われる．また，教師データとして Web から取得してきたブログの中に不適切なものが多く含まれていた可能性も考慮する必要がある．

表１　実験結果

評価項目	提案手法	従来手法
判定確率(%)	64.8	62.2
擬陽性率(%)	11.8	15.5
擬陰性率(%)	13.5	14.6

４．おわりに

本研究では，カテゴリ別にスパム判定コーパスを作成し，ブログ記事の記述内容を考慮したスパム判定手法を実現した．実証実験では，３つの評価項目において従来手法よりも優れた結果を出しており，本手法がブログスパムの判定に有効であることを実証した．しかし，今後の課題として，教師データが不適切な場合や，コメントやトラックバック元の記事の文字量が少ない場合，カテゴリを誤判定するため，適切なスパム判定コーパスを選択できない問題がある．これを解決するために，各単語の共起情報やブログ運営者と投稿者間の関係情報などの判定に有用なメタ情報を追加することで，文字量が少ない場合にも対応できるように研究を発展させる予定である．

参考文献

[1]Deniss Fetterly，Mark Manasse，Marc Najork：Spam，Damn Spam，and Statistics；Using Statistical Analysis to Locate Spam Web Pages，Proceedings of the 7th International Workshop on the Web and Databases，Association for Computing Machinery，pp.1-6，2004.6.

[2]池田大輔，山田泰寛，田中省作，松本英樹：部分文字列の数え上げによるブログスパムの検出，データベースシステム研究会研究報告，情報処理学会，Vol.2006，No.59，pp.45-52，2006.5.

[3]岩永学，田端利宏，櫻井幸一：ベイジアンフィルタリングを用いた迷惑メール対策における多言語環境でのコーパス分離手法の提案と評価，情報処理学会論文誌，情報処理学会，Vol.46，No.8，pp.1959-1966，2005.8.

[4]長谷川和男，西園敏弘，竹中豊文：文章自動分類における名詞を用いた分野の特徴選択，オフィスインフォメーションシステム研究会技術研究報告，電子情報通信学会，Vol.103，No.77，pp.19-24，2004.3.

[5]Gary Robinson：A Statistical Approach to the Spam Problem，Linux Journal，Specialized Systems Consultants，Vol.107，pp.58-64，2003.3.