様式 C-19

科学研究費補助金研究成果報告書

平成２２年３月３１日現在

研究種目：若手研究（B）
研究期間：2008～2009
課題番号：20700127
研究課題名（和文）スパムブログ空間の定量分析とフィルタリング手法の開発

研究課題名（英文）A study of spam blog filtering method based on its quantitative analysis

研究代表者
福原 知宏（TOMOHIRO FUKUHARA）
東京大学・人工物工学研究センター・特任助教
研究者番号：50436581

研究成果の概要（和文）：本研究ではWeb上のスパムブログ(splog)空間の定量分析に基づき，効率的にsplogをフィルタリングする手法の開発と評価を行った．日本語，中国語，英語，韓国語の splog を収集し，正解データ集合を作成した．正解データ集合を分析した結果，splog判定に個人差が見られることを確認した.機械学習を用いて各個人のsplog判定傾向を学習し，各個人に対して最適なsplogフィルタを提供するシステムの開発と評価を行った.日本語splog正解データ集合を用いて評価を行った結果，従来手法に比べ性能改善が可能であることを確認した.

研究成果の概要（英文）：A study of spam blog (splog) filtering method is conducted. For designing an efficient splog filter, we first created a splog dataset in which 50 persons judged blogs whether splogs or not. We then created a prototype splog filtering system. The system provides a user personalized splog filter by using a machine learning method called support vector machine (SVM). As evaluation results, we obtained F-value 0.738 for splogs which is higher than the value 0.656 of the previous method.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,500,000	450,000	1,950,000
２００９年度	100,000	30,000	130,000
年度			
年度			
年度			
総　計	1,600,000	480,000	2,080,000

研究分野：　総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：スパムフィルタリング

１．研究開始当初の背景

今日，Web 上の情報発信ツールとしてブログが普及している．多くの人々がブログサイトを開設し，日々の出来事や考えをブログ上で発信している．これらのブログは執筆者の備忘録や友人同士での閲覧だけでなく，Web検索の結果として利用されたり，商品やサービスに対するマーケティング調査に利用されており，利用価値は高い．

一方，近年は商品の宣伝や販売，商品販売サイトへの誘導を目的としたスパムブログ（splog）が大量に発生しており，Web 検索やブログを用いたマーケティング調査において問題となっている．これらの splog はどの言語においても存在することから，splog を効率的に識別し，フィルタリングする手法の開発が必要である．

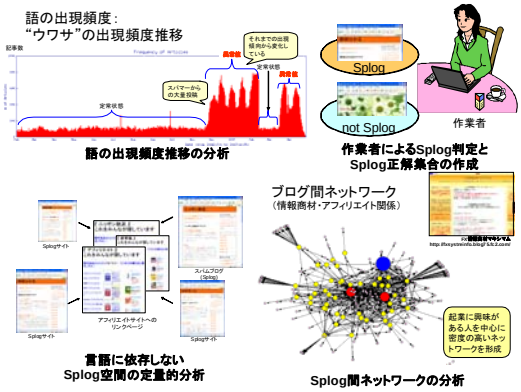

図 1．本研究の概要

２．研究の目的

本研究の目的は，splog の定量分析を通じ，splog を効率的にフィルタリングするシステムの開発である．本研究では日中韓英（日本語，英語，韓国語，中国語）の 4 言語における splog 正解集合を作成し，それぞれの splog 空間における語の出現頻度推移や splog 間ネットワークの分析を通じて splog 空間の特徴を定量的に記述し，言語に依存しない splog フィルタリング手法の研究開発を行う．図 1 に本研究の概要を示す．

３．研究の方法

（1）正解データ集合の作成

Splog フィルタリングシステムの開発を行うにあたり，splog 判定における正解データ集合を作成した．正解データ集合として，日本語 splog の正解データ集合と，日中韓英 4 言語の多言語 splog 正解データ集合を作成した．

①日本語 splog の正解データ集合では，日本人判定者 50 名による日本語ブログ記事 50 件に対する判定と，スパムブログ判定に関するアンケート調査を実施した．50 件のブログ記事は筆者らが事前に splog と判定したブログである．判定にあたっては，40 件の共通判定記事と 10 件の自由判定記事とに分け，前者は全ての判定者が判定する共通データとし，後者の 10 件は各判定者が自由に選択判定するデータとした．

判定では各記事に対し，判定者にとってのスパム度合いを示す spam 軸と，判定者にとって判定記事の情報にどれだけの価値があるかを示す value 軸とで判定を行った．

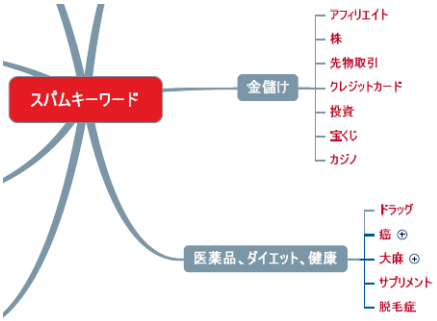

図 2．スパムキーワードの例

②多言語 splog データセットでは splog に出現するキーワードのリストを作成し，このキーワードを各言語に翻訳し，各言語の splog を収集した．キーワードのカテゴリとして，「アフィリエイト」「俳優・女優・タレント」「ビジネス」「ギャンブル」「アダルト」「金儲け」「医薬品・ダイエット・健康」の 7 カテゴリを用意し，それぞれのカテゴリについて数個のキーワードを用意した．図 2 にキーワードの一部を示す．

(2)Splog フィルタリングシステムの試作

（1）で作成した正解データ集合を用い，splog フィルタリングシステムの開発と評価を行う．正解データ集合に機械学習を適用し，splog フィルタリングの精度向上を目指す．

４．研究成果

(1) 正解データ集合の作成

Splog フィルタリングシステム開発において必要となる splog 正解データ集合の作成を行った．３．で述べた日本語 splog 正解データ集合と，多言語 splog 正解データ集合を作成した．

(2) Splog 判定における個人差の発見

日本語 splog 正解データ集合の作成と分析を通じて，splog 判定に個人差があることを発見した．図 3 に結果を示す．図 3 では，spam 軸の値が大きいものほどスパムを，value 軸の値が小さいものほど有益であることを示している．多くは価値が低くスパムと判定されているが，中には spam 値が高く value 値が低い splog も存在することが分かった．

図 4 に自由判定記事に対する判定結果を示す．自由判定記事の方が spam 値は低く，value 値もばらけることを確認した．

以上の結果から，splog フィルタリングでは個人に判定傾向に応じてフィルタを提供する必要があることを確認した．

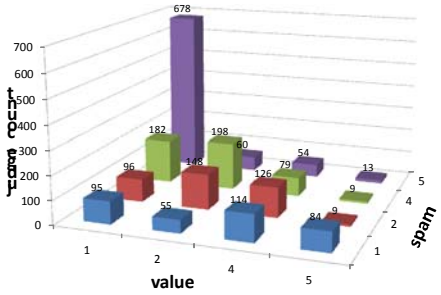

図 3. 共通記事 40 件に対する判定結果

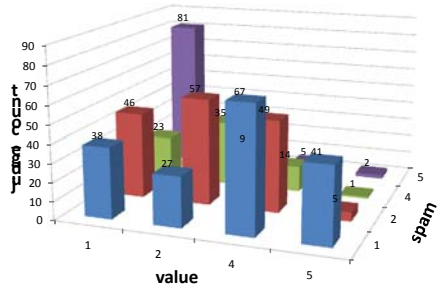

図 4. 自由選択記事 10 件に対する判定結果

(3) 個人適応型 splog フィルタリングシステムの開発と評価

機械学習を用いて各利用者の splog 判定傾向を学習し，各利用者に対して最適な splog フィルタを提供するシステムの開発と評価を行った．個人適応を行う場合と行わない場合での splog に対するフィルタリング性能を比較した．

日本語 splog 正解データ集合を用いて個人適応を行った場合，フィルタの性能を示す指標である F 値で平均 0.738，個人適応を行わない場合は平均 F 値 0.656 となり，個人適応によって 0.349 ポイントの改善が可能であることを確認した．また，個人適応により，最大で 0.718 ポイントの改善が可能であることを確認した．図 5 に結果を示す．

先行研究で報告されている手法との比較を行ったところ，提案手法により F 値の改善が可能であることを確認した．先行研究における平均 F 値が 0.656 であるのに対し，本研究での平均 F 値は 0.738 であり，0.082 ポイントの改善が可能であることを確認した（図 6 参照）．

本研究では日本語 splog に対するフィルタリングシステムの開発と評価を行ったが，当初予定していた多言語版の splog フィルタリングシステムの開発と評価までには至らなかった．今後，提案手法の多言語 splog への適用について調査を行う．

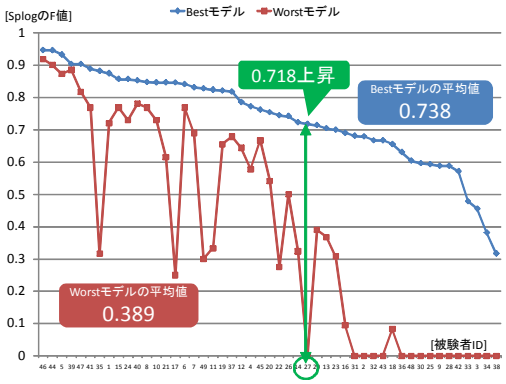

図 5. 個人適応を行う場合と行わない場合での splog に対する F 値の比較

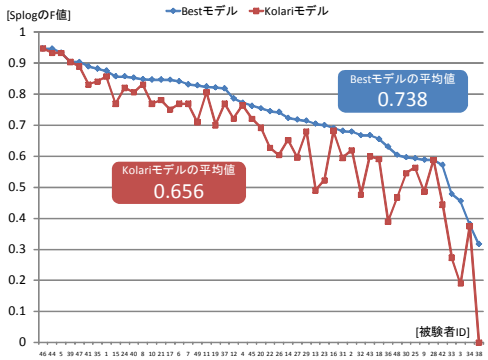

図 6. Splog に対する平均 F 値の比較（先行研究との比較）

(4) アフィリエイト ID を用いた splog 分析

splog の分析を進める中で，ブログサイト上に商品広告を掲載し，掲載された商品の売り上げに対して成功報酬を掲載者に提供するサービス（アフィリエイトサービス）が splog フィルタリングにおいて有用であることを確認した．

アフィリエイトサービスを利用しているブログ（アフィリエイトブログ）の記事を収集し，掲載広告数の順に上位 100 個のアフィリエイト ID のスパム判定を行った結果，56 個の ID がスパムと判定された．また，アフィリエイトブログにおける splog を調査したところ，splog では同一のアフィリエイト ID を使って複数のブログサイトが運営されていることを確認した．

アフィリエイトブログに対する解析システムを開発した（図 7 参照）．本システムにより Web ブラウザ上でどのようなアフィリエイトサイトがどのような商品を掲載しているかを把握できる．

今後，本システムを拡張し，アフィリエイトサービスを利用している splog に出現する商品や商品ジャンル等の情報を用いた splog フィルタリングについて研究を進める．

図 7. アフィリエイトブログ解析システム

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計 12 件）

① 芳中隆幸，福原知宏，増田英孝，中川裕志，機械学習を用いた個人適応型 splog フィルタリングの開発，日本データベース学会第 2 回データ工学と情報マネジメントに関するフォーラム,2010 年 3 月 2 日，淡路夢舞台国際会議場（兵庫県）．

② 石井聡一，福原知宏，増田英孝，中川裕志，ブログ上の広告活動を対象としたアフィリエイト分析支援システム，日本データベース学会第 2 回データ工学と情報マネジメントに関するフォーラム,2010 年 3 月 1 日，淡路夢舞台国際会議場（兵庫県）．

③ 芳中隆幸，福原知宏，増田英孝，中川裕志，個人適応型 splog フィルタリングシステムの実現に向けて: splog 判定データセットの構築と機械学習を用いたシステムの実装，電子情報通信学会言語理解とコミュニケーション研究会第 1 回集合知シンポジウム, 2010 年 1 月 25 日，広島市まちづくり市民交流プラザ（広島県）．

④ Katayama,T., Yoshinaka,T., Utsuro,T., Kawada,Y., Fukuhara,T., Detecting splogs using Similarities of splog HTML Structures, The 4th International Conference on Ubiquitous Information Management and Communication, 2010 年 1 月 14 日, Suwon (Korea).

⑤ 石井聡一，福原知宏，増田英孝，中川裕志，ブログ上の広告活動を対象としたアフィリエイト分析支援システム，日本データベース学会第 2 回データ工学と情報マネジメントに関するフォーラム，2009 年 11 月 21 日，慶應義塾大学日吉キャンパス（神奈川県）．

⑥ 石井聡一, 芳中隆幸，福原知宏，増田英孝，中川裕志, Web 上の広告情報を用いたアフィリエイトスパムの分析，楽天研究開発シンポジウム 2009，2009 年 11 月 14 日，品川シーサイド楽天タワー（東京都）．

⑦ 片山太一，宇津呂武仁，芳中隆幸，河田容英，福原知宏，HTML 構造の類似性を利用したスプログ検出方式，言語処理学会 NLP 若手の会 第 4 回シンポジウム，2009 年 10 月 1 日，京都大学百周年時計台記念館（京都府）．

⑧ Yoshinaka,T., Ishii,S., Fukuhara,T., Masuda,H., and Nakagawa,H.: A User-Oriented splog Filtering Based on Machine Learning, The 6th International Conference on Social Software (BlogTalk2009), 2009 年 9 月 15 日, Jeju (Korea).

⑨ 芳中隆幸，石井聡一，福原知宏，増田英孝，中川裕志，機械学習を用いたユーザ適応型 splog フィルタリングシステムの開発，第 23 回人工知能学会全国大会，2009 年 6 月 18 日，サンポートホール高松（香川県）．

⑩ 石井聡一，芳中隆幸，福原知宏，増田英孝，中川裕志，Web 上の広告活動の分析，第 23 回人工知能学会全国大会, 2009 年 6 月 19 日，サンポートホール高松（香川県）．

⑪ Katayama,T., Utsuro,T., Sato,Y., Yoshinaka,T., Kawada,Y., Fukuhara,T., An Empirical Study on Selective Sampling in Active Learning for splog Detection, The 5th International Workshop on Adversarial information Retrieval on the Web, 2009 年 4 月 21 日, Madrid (Spain).

〔図書〕（計 1 件）

① Yoshinaka,T., Ishii,S., Fukuhara,T., Masuda,H., and Nakagawa,H., Springer, Proceedings of the 6th International Conference on Social Software (BlogTalk2009) (Title: A User-Oriented Splog Filtering Based on Machine Learning), (in printing).

６．研究組織

(1)研究代表者

福原　知宏（TOMOHIRO FUKUHARA）
東京大学・人工物工学研究センター・
特任助教
研究者番号：50436581

(2)研究分担者

(3)連携研究者