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

A user-oriented Splog filtering system based on a machine learning method

芳中 隆幸[*1] 石井 聡一[*1] 福原 知宏[*2] 増田 英孝[*1] 中川 裕志[*3]
Takayuki Yoshinaka Soichi Ishi Tomohiro Fukuhara Hidetaka Masuda Hiroshi Nakagawa

[*1] 東京電機大学未来科学部
School of Science and Technology for Future Life, Tokyo Denki University

[*2] 東京大学人工物工学研究センター
Research into Artifacts Center for Engineering, The University of Tokyo

[*3] 東京大学情報基盤センター
Information Technology Center, The University of Tokyo

In this paper, we developed a user-oriented Splog filtering system using a machine learning technique. In the spam filtering domain, a personalized Splog filtering is important. We describe a user-oriented Splog filtering system that groups similar users into clusters, and provides Splog filters for each cluster. Our system uses the support vector machine (SVM) to classify Splogs in each cluster by learning Splog data in each cluster. We collected the training data for SVM from the assessment experiment, create two patterns and show that proposed method is effective. Moreover, developing a user-oriented Splog filtering system needs huge costs for using machine learning to each users. Therefore, we developed a Splog filtering system that classifies users into clusters.

1. はじめに

今日, インターネットへの情報発信の手段の一つとしてブログ (Weblog, ウェブログ) が普及し, 多くの人々が簡単にブログサイトを開設し, 情報発信できるようになった. その一方で, ブログサイトの中には, スプログ (Splog) と呼ばれるブログスパムが発生し, 検索エンジンにおける不当な順位操作や検索結果における精度低下の原因となっている. また, Splog は日々進化を続けており, Splog の定義 [Kolari 07] は存在するが, その認知は低い傾向にある [芳中 09]. ブログはその特性上ユーザの興味により必要とする情報の判断が異なる傾向にある [芳中 09]. そのため効果的な Splog フィルタリングには個々のユーザの興味を反映できるような柔軟なフィルタリングが必要である. そこで, 本論文ではユーザ適応型の Splog フィルタリングの開発を目指し, 被験者を用いた評価実験を行い, その結果を元に機械学習を用いたフィルタの開発を行った.

以下, 本論文の構成は次の通りである. 2. では, 現状における Splog 研究のレビューを行う. 3. では, 被験者を用いた評価実験についての分析結果とその考察について述べる. 4. では, 3. で得られたデータを元に機械学習を用いたユーザ適応型 Splog フィルタの評価を行う. 5. では, 本論文のまとめと今後の課題について述べる.

2. Splog フィルタリングの現状

Kolari は英語圏の Splog 空間に対して調査を行い, Splog 検知手法として SVM(Support Vector Machine)[Drucker 99] を利用することで F 値約 70%の Splog 検知に成功している [Kolari 06]. 日本語圏における Splog 研究としては石田 [石田 08] の研究があり, リンク解析に着目した研究を行い, F 値 90%以上の Splog 検知に成功している. 我々はこのような Splog フィルタリングはユーザ間で共通に存在する 1 つの Splog フィルタリングだと考えている.

一方, 我々が提案するユーザ適応型 Splog フィルタリングは, ユーザの興味に最適化されたフィルタリングを提供するため, ユーザは真に必要とする情報だけを取得することができる. また, 本フィルタリングの実現にあたり, 我々は 2 つの Splog 空間 (1) 万人に共通する Splog 空間と (2) ユーザごとに異なる Splog 空間の 2 種類が存在すると考えている [芳中 09]. このため, Splog フィルタリングには, ほぼ普遍的でありかつ随時更新可能な共通フィルタリング部とユーザごとに異なる個人適応型フィルタリング部の 2 つが有効だと我々は考える. 本論文では, (2) 個人適応型フィルタの開発を主として進めていく. また, ユーザ適応型 Splog フィルタリングの開発に必要となる, 正解情報は個々のユーザにおける判定情報をその正解情報として扱う.

我々は被験者を用いた評価実験から被験者のクラスタリングを用いたユーザ適応型 Splog フィルタリングの応用も試みる. ユーザ適応では, ユーザごとの Splog フィルタ作成が必要であり, 学習時に必要となる計算コストや, Splog フィルタ数などのコストが膨大になることが予想される. 本研究ではユーザ集合を複数のクラスタに分類し, 各クラスタに対してフィルタを提供することで, 個々のユーザにフィルタを提供する場合に比べ, フィルタ作成に要する計算コストを抑えることができると考えた. そこで本論文ではユーザのクラスタリングを用いたユーザ適応型の Splog フィルタリングへの応用とその評価も行う.

3. Splog 判定データの作成

クラスタリングを用いたユーザ適応型 Splog フィルタの開発に当たり, Splog に関する被験者実験を通じて Splog データセットを作成した. 本節では, 被験者実験における実験結果の

連絡先: 芳中 隆幸, 東京電機大学未来科学部情報メディア学科, 〒 101-8457 東京都千代田区神田錦町 2-2, 03-5280-3281, yoshinaka@cdl.im.dendai.ac.jp

分析とその考察を行う．また，本評価実験から，ユーザ適応型 Splog フィルタリングの開発に必要なデータの収集も行う．

3.1 Splog 判定評価実験概要

本評価実験は，評価実験用システム (SplogChecker システム) [芳中 09] を利用した Splog 判定評価実験を行った．

3.1.1 被験者

計 50 名で男女比 1:1(男性 25 名，女性 25 名) となるように被験者を募った．被験者集団の年齢は 21 歳から 55 歳まで，職業はコンピュータ関連技術，事務職，サービス職が占めている．被験者は普段からインターネットを利用している人を対象とした．

3.1.2 テスト記事

1 被験者に当たり，50 件のテスト記事を設ける．被験者間で共通に判定を行う記事を 40 件 (**共通記事**)，被験者間で個別に判定を行う記事を 10 件 (**個別集合**) 用意した．これらの記事は我々が選定を行い，一概にスパムとは判断しかねるような記事を選定した．これらの記事は 2009 年 1 月から 3 月までに筆者らが収集した記事群より選定を行った．

3.1.3 実験方法

被験者 50 名が SplogChecker システムを利用してテスト記事に対する Splog/ 非 Splog の判定を行う．本論文では，単純にスパムか非スパムかの 1 次元 2 値の判定基準ではなく，2 次元 4 値の判定基準を採用する．すなわち，1 次元目にはスパム軸 (*spam*) を採用し，2 次元目には情報としての価値軸 (*value*) を採用する．*spam* は記事のスパム度を表しており，「5 スパムである」「4 どちらかと言えばスパムである」「2 どちらかと言えばスパムでない」「1 スパムでない」の 4 段階評価で行う．*value* は記事の有用性を表しており，「1 有益であった」「2 どちらかといえば有益」「4 どちらかと言えば有益でない」「5 有益でない」の 4 段階評価で行う．また 4 値には，それぞれ「5, 4, 2, 1」の値を割り当ててある．また，判定にはタスクを設ける．共通記事 40 件にはキーワードが割り当てあり，被験者は割り当てられたキーワードで検索を行った結果の記事がテスト記事である場合の判定というタスクの下で判定を行う．個別記事 10 件は被験者の「最も興味のある」ジャンルと「最も興味のない」 ジャンルを 12 個のジャンル*1の中から選択し，被験者が記事を選定することで判定を行うというタスクを設ける．

3.2 実験結果

本評価実験では，ユーザ適応型 Splog フィルタの開発に必要となる重要なデータを収集した．以下，評価結果から得られたデータと分析結果について報告し，考察する．

第 1 に，共通記事 40 件において，それぞれの記事の判定情報 (*spam*, *value*) において，全被験者からの平均値を算出し 2 次元の分布図を作成した．図 1 にその分布を示す．図 1 において，2 次元 4 値から 4 分割されたグラフの右上に記事が多く分布していることがわかる．また，図 1 中，*spam*=3.5, *value*=4 に分布しているプロットがこの分布の平均値である (図 1 中，丸で囲まれた部分)．

第 2 に，図 2 に共通記事 40 件における全判定結果 (40 記事 × 50 被験者=2, 000 判定) の分布を示す．図 2 は，3 次元のグラフ軸を取り，奥行き軸が *spam*，横軸が *value* ，縦軸が「判定数」となっている．図 2 において，判定数 678 を示した分布があるのが，これは *spam*=5, *value*=5 を示しており，これらは一般に不要と扱われるノイズサイトであると言える．対して図 2 中，丸で囲まれた部分は，*spam* が高く，*value* が低いところを表している．つまり，スパムに分類される記事でも価値のあるブログサイトが多く存在していることがわかる．図 1 と比較すると，判定の平均では *spam* が高く，*value* が低い記事が分布に表れなかったが，全判定結果の分布では，そのような記事の存在がわかる．

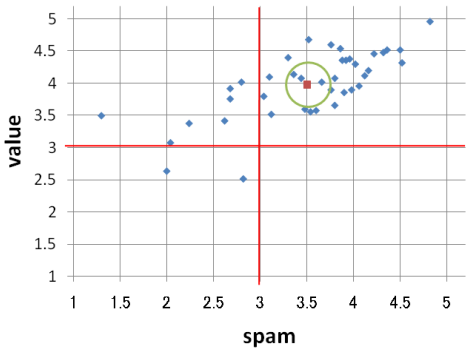

図 1: 2 次元 4 値の判定情報の平均を元にした共通記事 40 件の分布 (図中，丸で囲まれた部分は，分布の平均値である)

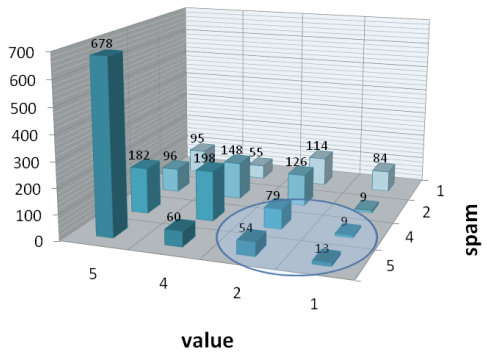

図 2: 被験者共通記事における 2 次元 4 値における判定結果分布

第 3 に，図 3 に被験者個別に選択可能な個別記事 10 件の判定結果の分布図を示す．図 3 の軸の取りかたは，図 2 と同様である．個別記事は，1 被験者に対して 10 件なので，併せて 500 件の判定情報が収集できた．図 3 は，図 2 に比べ *spam* は全体的に低い傾向にあることがわかる．これは，3.1.3 で説明したように，個別記事の選択には，「最も興味がある」記事と「最も興味がない」記事というタスクを設けているため，興味がある記事に対しては，スパムでないと判定する傾向にあることがわかる．

また，本評価実験では，被験者が判定に利用する判定特徴 [芳中 09] の収集も行った．判定特徴とは，被験者が判定時に利用した，ブログ記事内に存在する特徴である．表 1 にその集計結果を示す．特徴の使用頻度は 3 段階に設定し，「◎」は最も頻度があった特徴，「○」は一度でも利用した特徴，「×」は最も頻度が少なかった特徴と設定した．表 1 において，最も重要視されている特徴は「a. ブログ記事本文」であることがわかる．また，「f. 記事へのコメント」「g. トラックバック」「h. 記事本文の外にある広告」は判定における特徴としての利用頻度が低いことがわかる．そして，表 1 の集計結果から被験者のクラスタリングを行った．図 4 に表 1 から算出した階層クラス

*1 http://dir.yahoo.co.jp

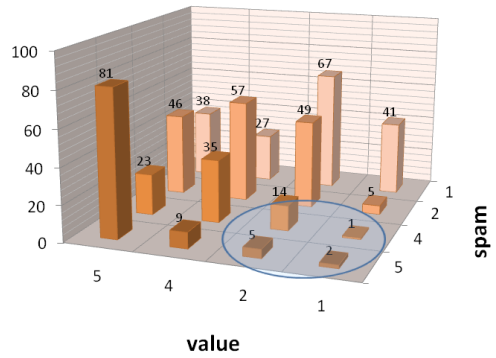

図 3: 被験者個別記事における 2 次元 4 値における判定結果分布

表 1: 判定特徴の集計結果 (数字は人数)

	◎	○	×
a ブログサイトのタイトル	5	35	9
b ブログ記事タイトル	8	37	4
c ブログ記事本文	33	15	1
d 記事本文内の画像	5	33	10
e リンク内のテキスト	7	29	13
f 記事へのコメント	0	10	39
g トラックバック	0	10	40
h 記事本文の外にある広告	0	23	26

タリングによるクラスタリング結果を示す. クラスタリングには統計解析ツール「R*2」を用い, クラスタ間の距離の算出には Ward 法を用いた. 図 4 のクラスタリング結果では, 距離 5 (図 4 中赤線の部分) で切った場合, 大きく 5 つのクラスタで形成されていることがわかる.

3.3 評価実験考察

本論文では, 判定の評価軸として 2 次元 4 値という判定基準を採用した. システム利用における評価実験では, 図 1 のように *spam* が高く, *value* が高いという記事群が多く見られた. これは一般的なノイズ情報であり, ユーザ間で共通に扱うフィルタ必要であるとわかる. また, 図 2 は, 全判定情報における分布を算出したもので, *spam* が高く, *value* が低い記事が存在していることがわかる. 図 1 と比べるとその点で大きく異なっている. *spam* が高く, *value* が低いような記事は全体として少ないため, このような記事に対してはユーザ適応型のフィルタが適当であると考えられる. 図 3 は個別記事による判定結果の分布であるが, 図 3 は図 2 に比べ全体的に *spam* が低くなっていることがわかる. これは被験者の興味が判定に影響していると考えられる. 図 4 において, 距離 5 からのクラスタの形成が見られた. これは, クラスリングを行うことによる, ユーザ適応型 Splog フィルタの可能性を示唆している. クラスタリングを行うことで, 学習時における計算コストの削減や, 必要となる Splog フィルタ数の冗長を省くことが可能となる.

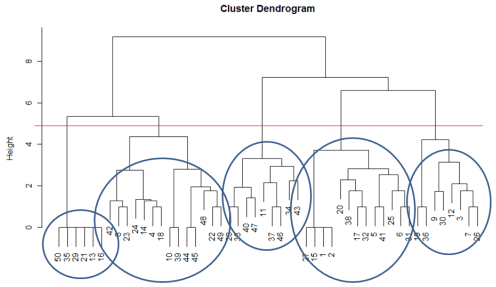

図 4: 判定特徴を利用した被験者クラスタリング

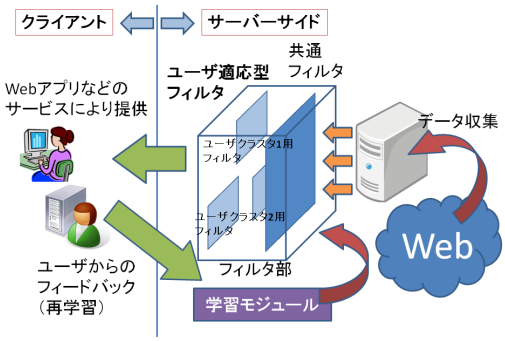

図 5: ユーザ適応型 Splog フィルタシステムの構成

4. ユーザ適応型 Splog フィルタシステム

本評価実験で得られたデータを元にユーザ適応型 Splog フィルタリングの開発を行った. 開発したユーザ適応型 Splog フィルタリングの構成を図 5 に示す. 開発したユーザ適応型 Splog フィルタリングは, サーバサイド側で動作する構成となっている. サーバ側で分類器として動作することで, ブログデータを分類し, それを Web アプリケーションなどのサービスを用いることでユーザ側へ提供する形となっている. 本フィルタは機械学習機能を持つ. 本論文では, 学習ツールとして LibSVM*3を用いる.

4.1 学習データ作成

学習データの作成には, 本評価実験おける, 被験者それぞれの判定データを正解情報として学習データの作成を行う. 2 次元 4 値の判定情報を 0, 1 の 2 値に置き換える事で学習を行う. 1(スパム) となる条件は, 「*spam* > 3 かつ *value* > 3」の時とし, それ以外の条件における判定情報の場合には, 0(非スパム) として判定情報とする. 学習に使用する特徴は, 数値的特徴 [芳中 09] を使用する. 本論文で使用する数値的特徴は, 「アフィリエイト ID 数」「外部リンク数」「内部リンク数」「画像数」「キーワード数」「文字数」「タグ無し文字数」である. 特徴は全部で 7 種類で, それぞれブログ記事の HTML から抽出した特徴となっている. これらの特徴と判定情報を元に学習データの作成を行う. 本論文では, 2 タイプの学習データの作成を行う. (1)7 種全ての特徴を含んだ学習データを作成する. (2)7 種全ての組合せによる学習データを作成する. (2) における学習データの総数を表した式を式 (1) に示す. また, n は特徴数

*2 http://www.okada.jp.org/RWiki

*3 http://www.csie.ntu.edu.tw/~cjlin/libsvm/

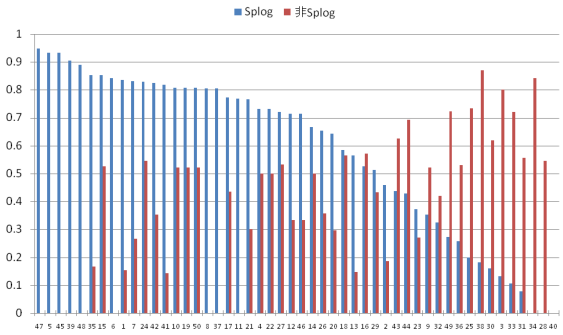

図 6: 全特徴による被験者ごとの学習結果

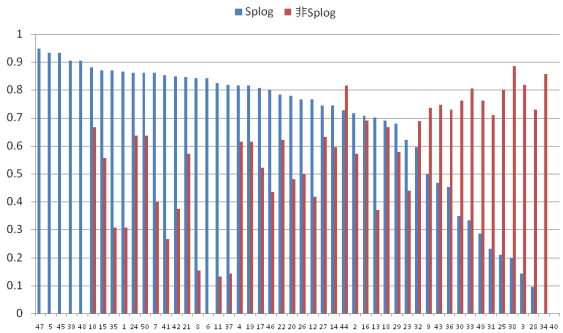

図 7: 特徴の組合せによる被験者ごとの学習結果

である.

$$total = \sum_{r=1}^{n} {}_nC_r \qquad (1)$$

本論文内で使用している 7 種類特徴の場合 $n = 7$ となり上記の式より, 6350 通りの学習データが作成されることになる.

4.2 評価結果

LibSVM を用いた 5 分割交差検定により評価を行う. 評価尺度には「F 値」を用いる.

第 1 に, 表 6 に (1) 全ての特徴を含んだ学習データによる学習結果を示す. 7 種全特徴による被験者ごとの学習結果において, Splog は, 被験者 47 の 0.947 が最高値で, 非 Splog は被験者 38 の 0.886 が最高値であった.

第 2 に, (2)7 種全ての組合せにより作成した学習データに対して学習を適応し, その評価結果中で, 最も F 値が良い数値となった結果を図 7 に示す. 組合せを用いた学習データの結果において, Splog は, 被験者 47 の 0.947 が最高値で, 非 Splog は被験者 38 の 0.870 が最高値であった.

第 3 に, 3.2 で算出した図 4 において, 各クラスタごとにおける評価を行う. 各クラスタで使用する特徴は, 7 種全ての特徴を使用し, クラスタの判定情報はクラスタ内における被験者群の判定情報の平均値から 3.1.3 の条件の下, 判定情報の算出を行い評価を行う. 図 8 にクラスタごとに行った学習結果を示す. 図 8 において Splog は, *cluster*2 の 0.486 が最高値で, 非 Splog は *cluster*3 の 0.974 が最高値であった.

4.3 考察

まず, 図 6 の 7 種全ての特徴を利用した被験者ごとにおける学習結果において被験者ごとの評価結果にばらつきが見ら

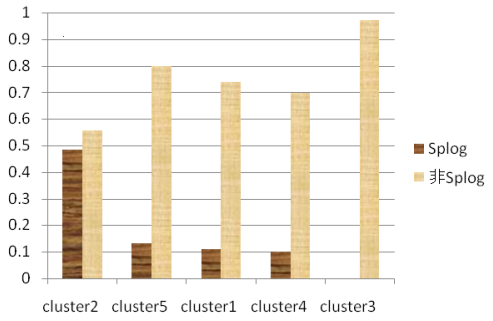

図 8: クラスタ別の分類性能 (Splog/非 Splog に対する F 値)

れることがわかる. これは, ユーザごとに効果のある特徴が異なって存在するということが考えられる. そこで我々は, 図 7 に示す 7 種全ての特徴の組合せを用いた学習データの作成を行い評価を行った. 図 7 の結果からもわかるように, 被験者ごとに最適な特徴の組合せは異なっており, ユーザそれぞれに最適な特徴を提供することが必要であることがわかる. 図 8 は, クラスタ形成後のクラスタごとにおける学習結果である. クラスタごとにおける学習結果でも, クラスタそれぞれの F 値に偏りがあることから, それぞのクラスタに対しても最適な特徴組合せを提供することで, クラスタリングを用いたユーザ適応型 Splog フィルタリングが十分に可能だと筆者らは考える.

5. おわりに

本論文では, 被験者を用いた評価実験から, 機械学習を用いたユーザ適応型 Splog フィルタリングに必要となるデータの収集を行い, ユーザ適応型 Splog フィルタリングの開発を行った. 評価には, 2 つのデータパターンを作成し評価を行うことでユーザそれぞれに最適化された特徴を提供することが必要であることを示した. また, 評価結果からクラスタリングによるユーザ適応型 Splog フィルタを作成し, 評価を行った結果, クラスタごとにも最適化された特徴を提供することでクラスタ適応型の Splog フィルタリングが有効であることを示した. 今後は, 学習結果の向上を目的とし, 特徴選定を行う他, 実際にフィルタを提供するためのサービス構築も行っていきたいと考えている. なお, 本研究は科研費 (20700127) の助成を受けたものである.

参考文献

[Drucker 99] Drucker, H., Wu, D., and Vapnik, V.: Supprt vector machines for Spam categorization, *IEEE-NN*, pp. 1048–1054 (1999)

[Kolari 06] Kolari, P., Finin, T., Java, A., and Joshi, A.: SVMs for the Blogosphere: Blog Identification and Splog Detection, *In Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, pp. 92–99 (2006)

[Kolari 07] Kolari, P., Finin, T., Java, A., and Joshi, A.: Towards Spam Detection at Ping Servers, *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)* (2007)

[石田 08] 石田 和成：共起クラスターシードと連鎖的抽出にもとづくスパムブログのフィルタリング, データベースと Web 情報システムに関するシンポジウム DBWeb2008 (2008), 2B-1

[芳中 09] 芳中 隆幸, 福原 知宏, 増田 英孝, 中川裕志：ブログ空間におけるスパムサイト解析ツールの開発-ユーザ適応型 Splog フィルタリングに向けて-, 暗号と情報セキュリティシンポジウム SCIS2009 (2009), 1E1-3