3B1-4

個人適応型Splogフィルタリングシステムの開発と評価

Development and evaluation of a user-oriented splog filtering system

芳中隆之[*1] 福原知宏[*2] 増田英孝[*1] 中川裕志[*3]
Takayuki Yoshinaka 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]独立行政法人産業技術総合研究所
National Institute of Advanced Industrial Science and Technology

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

A user-oriented Splog filtering system and its evaluation results are described. Proposed system provides a user the best splog filtering model by learning his/her evaluation data for splogs. The system finds the best model for the user by learning several sets of features and kernel functions. The system uses the support vector machie (SVM), and uses (1) eight feature sets and (2) four kernel functions. We obtained the average F-measure value 0.738 for splogs that is larger than the value obtained by using the previous work.

1. はじめに

近年，多くの個人や組織がブログ上で情報発信を行っている [総務 09]．これらのブログの中には有益な情報も含まれるが，スプログ (Splog) と呼ばれるスパムも存在し，Web 検索における問題となっている [Katayama 09]．ブログ記事は電子メールに見られるスパムのように明確にスパムか否かを判定できる訳ではなく，個人や状況によって判定が分かれるものも存在する [芳中 10]．

我々は個人の Splog 判定傾向を学習し，個人毎に最適なフィルタを提供する個人適応型 Splog フィルタリングを提案している [芳中 10]．このフィルタリングでは Support Vector Machine (SVM)[Vapnik 95] と特徴量の組合せを用いることで，個人毎に最適な Splog フィルタを作成する．

本論文では提案手法の評価として，50 名の個人別 Splog 判定データセットを用い，個人別の最適 Splog 判定モデルの算出と評価を行った．実験結果と個人適応型 Splog フィルタリングの有効性について述べる．

本論文の構成は次の通りである． 2. では個人適応型 Splog フィルタリングシステムについて述べる． 3. では評価実験について述べる． 4. では実験結果の考察を行う． 5. ではまとめと今後の展望について述べる．

2. 個人適応型 Splog フィルタリングシステム

本手法は既に筆者らが作成した個人別 Splog 判定データセット [芳中 10] を元に，SVM を用いて各個人の最適 Splog 判定モデルの構築を行う．図 1 に本手法の概要を示す．

個人別最適 Splog 判定モデルの構築にあたり，本研究では 8 種類の特徴群と 4 種類のカーネル関数 (線形カーネル，3 次多項式カーネル，RBF カーネル，シグモイドカーネル) の組合せを用いる．これにより各人に対して 32 種類の学習データを用意し，学習を行う．学習には LibSVM (version 2.88)[*1]を用いた．以下，splog 判定データセットと機械学習で用いる 8 種類の特徴群について説明する．

連絡先: 福原知宏, 独立行政法人産業技術総合研究所, 東京都江東区青海 2-3-26

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

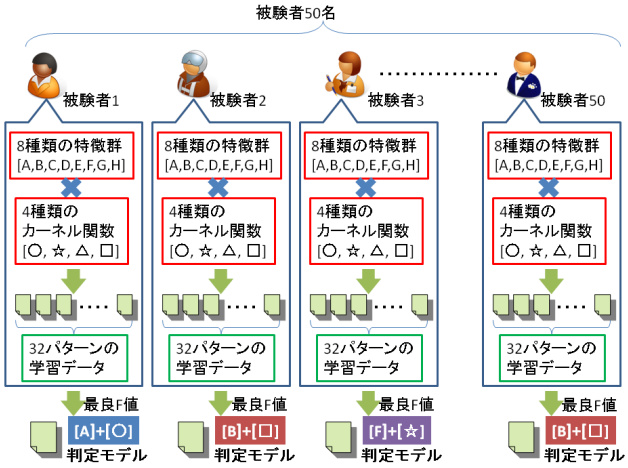

図 1: 個人別最適 Splog 判定モデルの算出方法

2.1 Splog 判定データセット

評価用データセットとして，被験者 50 名により作成した Splog 判定データセットを用いる [芳中 10]．このデータセットには被験者間で共通の 40 件のブログ記事に対する被験者ごとの splog 判定情報が含まれている．

2.2 8 種類の特徴群

表 1 に学習に用いる 8 種類の特徴群を示す．これらの特徴群は，筆者らが提案する特徴群 (Original features) と，先行研究 [Kolari 06] で示された特徴群 (Kolari features) とに分かれる．以下，各特徴群について述べる．

Original features

表 1 にある Original features は筆者らが提案する抽出が容易な特徴群 (Light-weight features; LWF) から構成されるカテゴリである．LWF はブログ記事の HTML テキストのみから抽出可能な特徴群であり，次元数は 12 である．以下，LWF に含まれる特徴群について述べる．

キーワード数

表 1: 8 種類の特徴群

カテゴリ名	ID	特徴群名	次元数
Original Features	1	Light-weight Features	12
Kolari features	2	Specialized features	13
	3	Bag-of-URLs	3,091
	4	Bag-of-anchors	4,014
	5	Bag-of-words	9,014
	6	Bag-of-Kolaris	16,119
	7	2-gram features	34,769
	8	3-gram features	47,555

ブログ記事の本文部分から抽出した名詞 (未知語を含む) の数.

句点の数
HTML タグを排除したブログ記事からカウントした句点 (「。」) の数.

読点の数
HTML タグを排除したブログ記事からカウントした読点 (「、」) の数.

ブログ記事全体の文字数
HTML タグを含むブログ記事全体の文字数.

(タグを除いた) ブログ全体の文字数
HTML タグを含まないブログ記事全体の文字数.

ブレイクタグの数

タグの数.

内部リンク数
同一ブログサイト内へのリンク数. 例えばブログ記事内におけるコメントやアーカイブへのリンクが該当する.

外部リンク数
同一ブログ記事以外へのリンク数.

画像ファイル数
ブログ記事に含まれる画像ファイル数.

画像サイズ (高さ) の平均値
ブログ記事中に出現する画像の高さサイズの平均値.

画像サイズ (幅) の平均値
ブログ記事中に出現する画像の幅サイズの平均値.

アフィリエイト識別番号数
HTML 内に出現するアフィリエイトサイトへのリンクから抽出したアフィリエイト識別番号の個数.

Kolari features

Kolari features とは Kolari らによる先行研究 [Kolari 06] で使用された特徴群であり,「単語」,「アンカーテキスト」,「URL」等の特徴量が含まれる. 本研究では個人適応型 Splog フィルタリングにおけるこれらの特徴群の効果を調べる.

表 1 から Kolari features は Specialized features, Bag-of-URLs, Bag-of-anchors, Bag-of-words, Bag-of-Kolaris, 2-gram features, 3-gram features の 7 つの特徴群から構成される. これらの抽出方法は Kolari らの抽出方法と同一である. 以下に Kolari features の各特徴群について述べる.

Specialized features
「単語」「アンカーテキスト」「URL」の情報を先行研究の方法で抽出した特徴群である. 次元数は 13 であり, 値には真偽値 (1 or 0) と tf*idf 値 [Manning 99] が用いられる.

Bag-of-URLs
ブログ記事中に出現する全 URL において,「.(ドット)」と「/(スラッシュ)」で分割した文字単位での URL 情報を使用する. また「http://」「www」は抽出対象外とする. 抽出された Bag-of-URLs の次元数は 3,091 であり, ベクトル値には真偽値を用いる.

Bag-of-anchors
ブログ記事中における<a>タグで囲まれた部分のテキスト情報において, 全ての品詞情報における形態素情報を抽出した特徴群である. 抽出された Bag-of-anchors の次元数は 4,014 であり, ベクトル値には真偽値を用いる.

Bag-of-words
ブログ記事中において全ての品詞情報における形態素情報を抽出した特徴群である. 抽出された Bag-of-words の次元数は 9,014 であり, ベクトル値には tf*idf 値を用いる.

Bag-of-Kolaris
Bag-of-URLs, Bag-of-anchors, Bag-of-words を複合した特徴群である. 次元数は 16,119 である.

2-gram features, 3-gram features
Bag-of-words を bi-gram と tri-gram により抽出した特徴群である. 次元数はそれぞれ 34,769, 47,555 であり, ベクトル値は真偽値を用いる.

なお, 表 1 の Kolari features における次元数は評価用データセットにおける次元数である.

3. 評価実験

評価方法と実験結果について述べる.

3.1 評価方法

被験者ごとに提供される 32 種類の学習データの評価尺度として, 各被験者の判定データを用いて Splog に対する F 値を用いた. この時, 5 分割交差検定を行い, F 値の平均値を用いた.

3.2 Splog 判定モデルの評価結果

図 2 に個人ごとの「F 値が最良となった特徴群とカーネルの組合せによる Splog の F 値 (Best モデル)」と,「F 値が最低値となった特徴群とカーネル関数の組合せによる Splog の F 値 (Worst モデル)」を示す. 図 2 中, y 軸は Splog に対する F 値であり, x 軸は被験者 ID である. 被験者 ID は Best モデルにおける Splog の F 値でソートされている.

図 2 における Best モデルは 32 種類の学習結果において最良の F 値を示した特徴群とカーネルの組合せである. つまり, Best モデルが各被験者に対する最適 Splog 判定モデルとなる. 最良 F 値を示した被験者は, 被験者 ID47 の 0.947 (図 2 中, 最も左の被験者) であった. Worst モデルは 32 種類の学習結果において最も低い F 値となったモデルであり, F 値が 0 となった被験者が 16 名存在した. それぞれのモデルにおける Splog の平均 F 値は Best モデルが 0.738 で, Worst モデルが 0.389 であった. それぞれの平均 F 値の差分は 0.342 であり, 全体での 0.3 以上の F 値向上が見られた. また, 被験者ごとに Best モデルと Worst モデルの F 値の差分を算出した場合, 被験者 ID27 の 0.718 (図 2 中, 丸で囲まれた被験者) が最大の F 値向上となった.

8 種類の特徴群の内, 各カーネル関数における Splog の平均 F 値が他の特徴群より良い値となったのは Bag-of-URLs であった. Bag-of-URLs の各カーネル関数における平均 F 値は線形カーネルが 0.656, 多項式カーネルが 0.703, RBF カーネルが 0.531, シグモイドカーネルが 0.522 であり, これらの平

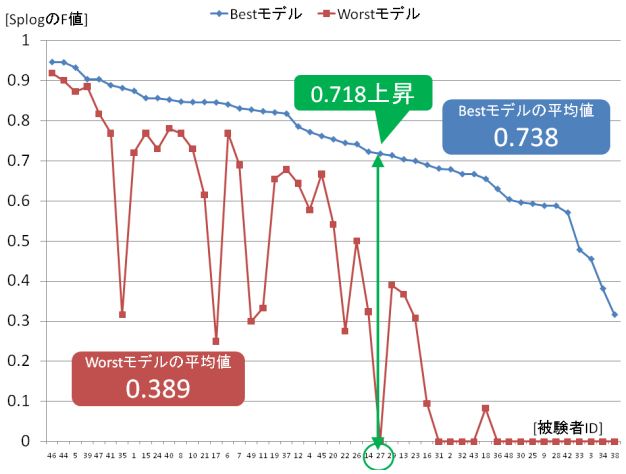

図 2: 各被験者の Best モデルと Worst モデルにおける Splog の F 値

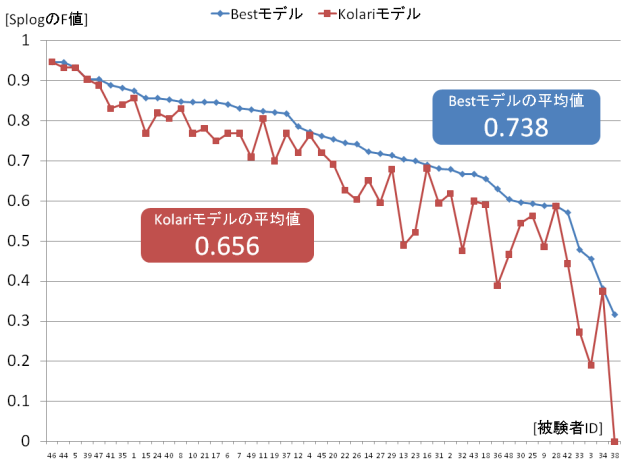

図 3: 各被験者の Best モデルと Kolari モデルにおける Splog の F 値

均 F 値は 0.652 であった．この平均 F 値が 8 種類の特徴群において最も良い値となったのが Bag-of-URLs であった．

また，Kolari らの調査では線形カーネルが有効だったと報告していることから，提案手法で算出した Best モデルと，Bag-of-URLs と線形カーネルを用いた各被験者の Splog 判定モデル (Kolari モデル) を比較した．図 3 に Best モデルと Kolari モデルにおける各被験者の Splog の F 値を示す．

図 3 において，それぞれのモデルにおける Splog の平均 F 値は Best モデルが 0.738 で，Kolari モデルが 0.656 であった．それぞれの平均 F 値の差分から，全体での約 0.08 の F 値向上が見られた．Kolari モデルにおける F 値が Best モデルにおける F 値と同値になることはあるが，Best モデルの F 値を越えることはなかった．

以上の結果から，先行研究で有効とされる特徴群とカーネル関数を用いた手法は個人適応型 Splog フィルタリングを考えた場合，必ずしも最適ではなく，本提案手法を用いることで，各ユーザに対して最適なフィルタを提供できることが分かった．

3.3 8 種類の特徴群の評価結果

特徴群 Original features と Kolari features の評価とカーネル関数の評価を行った．図 2 の結果において Best モデルで使用された特徴群とカーネル関数の出現頻度を集計した結果を表 2 に示す．

表 2 から，最も使用された特徴群は Light-weight features の 31 回であり，最も使用されなかった特徴群は Specialized features の 18 回であった．Light-weight features と Specialized features の出現回数においては 13 回の差が見られたが，Specialized features 以外の特徴群と比較した場合には，出現回数に大きな差は見られなかった．また，カーネル関数の出現回数の合計においても，線形カーネルが 46 回，多項式カーネルが 49 回，RBF カーネルが 55 回，シグモイドカーネルが 45 回と，大きな差は見られなかった．

これらの結果から各被験者に最適 Splog 判定モデルを決定するための最適な特徴群とカーネルの組合せが複数存在することが分かり，また，それらの出現回数には大きな差が生じないことが分かった．

4. 考察

表 2 の特徴群とカーネル関数の出現回数の結果から，各被験者には複数の最適な特徴群とカーネル関数の組合せが存在することが分かった．

ユーザ適応型 Splog フィルタリングにおいて問題となるのは計算コストである [Jeh 03]．我々は計算コスト削減のために，低次元の特徴群を用いた個人別最適 Splog 判定モデルの再算出を行った．これには各被験者が持つ複数の最適な特徴群とカーネル関数の組合せを表 1 における各特徴群の次元数を優先度とすることで，各被験者に対して一意の最適な特徴群とカーネル関数の組合せを決定した．これにより，例えば 3-gram features と Light-weight features で同一の Splog の F 値を算出した最適 Splog 判定モデルの場合，次元数の優先度を考慮することで Light-weight features を用いた最適 Splog 判定モデルの再算出が可能となり，計算コストの削減が可能となる．

表 3 に各特徴群の次元数における優先度を考慮した場合の特徴群とカーネル関数の出現回数をカウントした結果を示す．優先度を考慮した場合，最も出現頻度の多い特徴群は Light-weight features の 21 回であった．これは全体の 42% を占める出現回数である．Light-weight features と他の特徴群とを比較すると，表 2 に比べ，出現頻度に大きな差が生じていることが分かる．この結果，次元数に優先度を考慮した場合，個人適応型 Splog フィルタリングにおいては Light-weight features が有効だと言える．

一方，表 3 中，最も出現回数が少なかった特徴群は Bag-of-words の 2 回であった．Kolari らは Bag-of-words を使用した場合，約 80% の Splog 検出率を示しており，Splog フィルタリングではこの特徴群を使用することが有効だとしている．一方，個人適応型 Splog フィルタリングでは，Bag-of-words を用いることで有効な Splog 判定モデルが得られたのは 2 回だけであった．このことから，従来研究で示された高い Splog 検知率を示す特徴群でも個人適応型 Splog フィルタリングでは必ずしも最適ではないと言える．

以上の結果から，我々が提案する Original features と従来研究で用いられた Kolari features とを比較した場合，各特徴群の次元数を考慮した Original features (Light-weight features) が個人適応型 Splog フィルタリングにおいて有効だと分かった．

表 2: Best モデルに用いられた特徴群とカーネル関数の出現頻度

ID	特徴群名	出現頻度	線形	多項式	RBF	シグモイド
1	Light-weight Features	31	7	6	12	6
3	Bag-of-URLs	28	4	14	6	4
8	3-gram Features	25	7	6	7	5
7	2-gram Features	24	6	5	6	7
4	Bag-of-anchors	23	7	5	6	5
6	Bag-of-Kolaris	24	6	6	6	6
5	Bag-of-words	22	5	5	6	6
2	Specialized Features	18	4	2	6	6
合計		195	46	49	55	45

表 3: Best モデルに用いられた特徴群とカーネル関数の出現頻度 (優先度あり)

ID	特徴群名	出現頻度	線形	多項式	RBF	シグモイド
1	Light-weight Features	21	7	6	8	0
3	Bag-of-URLs	10	1	8	0	1
8	3-gram Features	5	2	0	1	2
2	Specialized Features	3	0	2	1	0
4	Bag-of-anchors	3	2	1	0	0
7	2-gram Features	3	1	1	0	1
5	Bag-of-words	2	1	1	0	0
6	Bag-of-Kolaris	2	1	1	0	0
合計		49	15	20	10	4

5. まとめ

本論文では個人適応型 Splog フィルタリングシステムについて述べ，提案手法の評価結果について述べた．先行研究との比較の結果，被験者全体で 0.35 の Splog の F 値向上が可能であり，個人別には最大 0.72 の Splog の F 値上昇が得られた．また，我々が提案する Light-weight features を先行研究の特徴群 (Kolari features) と比較することで，提案手法の有効性を示した．今後の課題として，個人別最適モデル算出に要する計算コストの削減，実装システムを用いたより多くの評価者による評価実験の実施について検討する．

参考文献

[Jeh 03] Jeh, G. and Widom, J.: Scaling personalized web search, *Proceedings of the 12th International Conference on World Wide Web (WWW2003)*, pp. 271–279 (2003)

[Katayama 09] Katayama, T., Sato, Y., Utsuro, T., Yoshinaka, T., Kawada, Y., and Fukuhara, T.: An Empirical Study on Selective Sampling in Active Learning for Splog Detection, *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb2009)*, pp. 29–36 (2009)

[Kolari 06] Kolari, P., Java, A., Finin, T., Oates, T., and Joshi, A.: Detecting Spam Blogs: A Machine Learning Approach, *Proceedings of the 21st National Conference on Association for Advancement of Artificial Intelligence (AAAI2006)*, pp. 1351–1356 (2006)

[Manning 99] Manning, C. D. and Shuetze, H.: *Foundations of Statistical Natural Language Processing*, MIT Press (1999)

[Vapnik 95] Vapnik, V. N.: *The nature of statistical learning theory*, Springer-Verlag New York, Inc. (1995)

[総務 09] 総務省情報通信政策研究所調査研究部：ブログ・SNS の経済効果に関する調査研究 (2009), (available at http://www.soumu.go.jp/iicp/chousakenkyu/seika/houkoku.html)

[芳中 10] 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, 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 (DEIM2010) (2010), B10-4