DEIM Forum 2009 B9-6

機械学習を用いたスパムブログ検出における信頼度の利用

片山　太一† 　佐藤　有記†† 　宇津呂武仁†† 　芳中　隆幸††† 　河田　容英††††
福原　知宏†††††

† 筑波大学第三学群工学システム学類　〒 305-8573 茨城県つくば市天王台 1-1-1
†† 筑波大学大学院システム情報工学研究科　〒 305-8573 茨城県つくば市天王台 1-1-1
††† 東京電機大学工学研究科　〒 101-8457 東京都千代田区神田錦町 2-2
†††† (株) ナビックス　〒 141-0031 東京都品川区西五反田 8-3-6
††††† 東京大学 人工物工学研究センター　〒 277-8568 千葉県柏市柏の葉 5-1-5

あらまし　本論文では，機械学習のひとつである SVM を用いた枠組みによって，スパムブログ (スプログ) の判定を行うタスクを設定する．未判定のブログが入力されたとき，SVM の分離平面によって，スパムブログかそうでないかを決定する．さらに，SVM では分離平面との距離を出力できるので，分離平面との距離を信頼度として利用する．出力である信頼度を用いて，ブログサイトを，高信頼度スパムブログ判定結果，高信頼度非スパムブログ判定結果，低信頼度判定結果の三種類に分ける．また，前後する 2 つの期間のデータセットを用いて，評価用データセットよりも以前のスプログを用いて訓練した分類器，および，評価用データセットと同時期のスプログを用いて訓練した分類器の間で性能を比較することにより，一定期間の間にスプログが変化していることを示す．このことより，時間の経過とともにスプログデータセットを更新して，新種のスプログに適切に対応した分類器を訓練する必要があることがわかる．そこで，少ない人手コストで的確に新種のスプログ候補を同定して，効率よくデータセットを更新することを目的として，能動学習の枠組みを適用した結果を報告する．さらに，同一の作成者が大量に作成した「大量生成型」のスプログの同定についても，評価実験を行う．
キーワード　スパムブログ，機械学習，信頼度，SVM

Estimating Confidence in Machine Learning for Splog Detection

Taichi KATAYAMA†, Yuuki SATO††, Takehito UTSURO††, Takayuki YOSHINAKA†††, Yasuhide KAWADA††††, and Tomohiro FUKUHARA†††††

† College Eng. Sys., Third Cluster of Colleges, University of Tsukuba,　Tsukuba, 305-8573, Japan
†† Grad. Sch. Systems and Information Engineering, University of Tsukuba,　Tsukuba, 305-8573, Japan
††† Graduate School of Engineering, Tokyo Denki University,　Tokyo, 101-8457, Japan
†††† Navix Co., Ltd.　8-3-6 Nishi-Gotanda, Shinagawa-Ku Tokyo 141-0031, Japan
††††† Research into Artifacts, Center for Engineering, University of Tokyo　Kashiwa, Chiba 277-8568, Japan

Abstract　This paper studies the issue of identifying spam blogs (splogs) by SVM. In an SVM based classifier, a separating hyperplane is used for classifying splogs and authentic blogs. In our study, we further utilize the distance from the separating hyperplane to an instance as a confidence measure. In our approach to semi-automatic collection of splogs and re-training of the SVM based classifier, we consider this confidence measure and identify blog sites which cannot be reliably judged whether splogs or authentic blogs. This paper further studies how to reduce the amount of human supervision for identifying splogs / authentic blogs in the context of continuously updating splog data sets year by year. We also examine whether we can automatically identify professional spammers who automatically create a number of similar splogs.
Key words　spam blog, machine learning, confidence measure, SVM

1. はじめに

ブログには個人の意見情報が記されており，市場の動向を推測するための手掛かりや製品についての意見調査をする上で有益であるとして，近年注目を集めている．そのため，従来からあるインデクシングのみを行う検索エンジンとは異なる，ブログ特有の情報検索サービスが出現している．

具体的には，ブログ解析サービスとして，*Technorati*(注1)，*BlogPulse*(注2)，*kizasi.jp*(注3)，*blogWatcher* (注4) [1] などが存在する．多言語ブログサービスとしては，*Globe of Blogs*(注5) が言語横断ブログ記事検索機能を提供している．また *Best Blogs in Asia Directory* (注6) がアジア言語ブログの検索機能を提供している．*Blogwise*(注7)もまた多言語ブログ記事の分析を行っている．

一方で，ブログのウェブコンテンツの作成と配信は非常に容易になっており，そのことが引き金となって，アフィリエイト収入を得ることを目的とするスパムブログ (以下，スプログ) が急増している [2]〜[6]．スプログにおいては，通常，広告主への誘導または対象サイトのページランクを増加する目的のもとで，機械的な文書作成や他サイトの引用という手段を用いて自動的に記事を生成し，大量のリンクを有するブログを機械的に自動生成する. [4] は英語ブログにおいて，約 88%のブログサイトがスプログであり，それは全ブログポストの 75%を占めると報告している．このことから, [3], [7] に述べられているように，スプログは情報検索品質の低下やネットワークと格納資源の多大な浪費などといった問題を起こす要因となる．そのため，近年，スプログの分析や検出を目的とした研究が進められている. [5] では，TREC(注8)Blog06 データコレクションを用いて，スプログのピング時系列特性，入力度数/出力度数の分布特性，典型的な単語群を分析している．また, [4], [6] は，*BlogPulse* データセットを用いたスプログ分析の結果を報告している．一方, [4], [8]〜[10] では，スプログを機械的に特定し，排除する技術について報告している．

以上の先行研究をふまえて，本論文では，機械学習のひとつである Support Vector Machines [11] (SVM) を用いた枠組みによって，スパムブログの判定を行うタスクを設定する．未判定のブログが入力されたとき，SVM の分離平面によって，スパムブログかそうでないかを決定する．さらに，SVM では分離平面との距離を出力できるので，分離平面との距離を信頼度として利用する [12]．出力である信頼度を用いて，ブログを，高信頼度スパムブログ判定結果，高信頼度非スパムブログ判定結果，低信頼度判定結果の三種類に分ける．

また，2007〜2008 年に収集したスプログ/非スプログデータセットと 2008〜2009 年に収集したスプログ/非スプログデータセットの 2 つのデータセットを用いて，評価用データセットよりも以前のスプログを用いて訓練した分類器，および，評価用データセットと同時期のスプログを用いて訓練した分類器の間で性能を比較することにより，一定期間の間にスプログが変化していることを示す．このことより，時間の経過とともにスプログデータセットを更新して，新種のスプログに適切に対応した分類器を訓練する必要があることがわかる．そこで，少ない人手コストで的確に新種のスプログ候補を同定して，効率よくデータセットを更新することを目的として，能動学習の枠組みを適用した結果を報告する．さらに，収集されたデータセットの中で，極めて構造が類似するスプログを，同一作成者が自動生成している「大量生成型」のスプログ [13], [14] として同定し，類似するスプログごとに ID をつけ，その ID 別に「大量生成型」のスプログを検出する実験を行った結果を示す．

(注1)：http://technorati.com/
(注2)：http://www.blogpulse.com/
(注3)：http://kizasi.jp/ (日本語のみ)
(注4)：http://blogwatcher.pi.titech.ac.jp/ (日本語のみ)
(注5)：http://www.globeofblogs.com/
(注6)：http://www.misohoni.com/bba/
(注7)：http://www.blogwise.com/
(注8)：http://trec.nist.gov/

2. スプログ/非スプログデータセット

本論文では，表 1 に示すように，2007 年 9 月 〜2008 年 2 月，および，2008 年 12 月 〜2009 年 1 月の 2 つの期間において収集した日本語スプログ/非スプログデータセットを用いる．いずれのデータセットにおいても, [13], [14] で提案された基準によって，スプログ/非スプログを判定した結果が付与されている．以下にその判定基準を示す．

（1） ブログサイトが以下のいずれかの条件を満たす場合，**スプログ**と判定する．

（a）「オリジナルの文章」が全く存在しない．

（b）「オリジナルの文章」はあるが,「アフィリエサイトへのリンクがある」「広告記事がある」「アダルトコンテンツを含む」のいずれかを満たす．

（2） それ以外の場合，そのブログサイトは**非スプログ**と判定される．

スプログ/非スプログの分布が 2 つの期間で異なるのは，ブログサイトの収集方法が異なるためである．2007〜2008 年に収集したデータセットにおいては，以下の条件を満たすブログサイトを対象として収集を行った．

i) スプログに頻出するキーワードを含む，複数のキーワードを選定し，各キーワードを含むブログサイトを対象とする．

ii) 事前の調査により，特定のキーワードに対して，そのキーワードを含むブログ記事数が最も多い日付であるバースト日に投稿されたブログ記事にスプログが多く含まれる傾向が高いことがわかっている．この傾向をふまえ，このバースト日に投稿されたブログ記事を含むブログサイトを対象とする．

iii) さらに，バースト日における更新頻度の高いブログサイトを優先的に収集対象とする．

一方，2008〜2009 年に収集したデータセットにおいては，より効率的にスプログが収集できるような方式を採用した. [13], [14] やその他の先行研究から，複数のスプログが同一のアフィリエイトサイトを共有する傾向があることがわかっている．この傾向をふまえ，この 2008〜2009 年に収集したデータセットにお

表 1　スプログ/非スプログデータセット中のスプログ/非スプログサイト数

(a) 全事例数

ブログ収集時期	スプログ数	非スプログ数	合計
2007〜2008 年	768	3318	4086
2008〜2009 年	1445	2459	3904
合計	2213	5777	7990

(b) 5.2 節の評価における事例数

2007〜2008 年	40×18=720	40×18=720	1440
2008〜2009 年	40×18=720 (40×10=400 のみ使用. 内，訓練事例 360 個，評価事例 40 個)	40×18=720 (40×10=400 のみ使用. 内，訓練事例 360 個，評価事例 40 個)	1440 (800 のみ使用. 内，訓練事例 720 個，評価事例 80 個)

(c) 5.3 節の評価における事例数

2008〜2009 年	訓練開始時，訓練データ 4 サイト, プール 1296 サイト，評価 145 サイト	訓練開始時，訓練データ 6 サイト，プール 2208 サイト，評価 245 サイト	訓練開始時，訓練データ 10 サイト，プール 3504 サイト，評価 390 サイト

表 2　能動学習における新規訓練事例の選択的サンプリング方式

		低信頼度側/高信頼度側のどちらから選ぶか?				
		高信頼度	低信頼度	高/低信頼度	バランス	ランダム
スプログ側/非スプログ側のどちら側から選ぶか?	スプログ	—	—	—	—	ランダム
	非スプログ	高信頼度-非スプログ	—	高/低信頼度-非スプログ	—	
	スプログ/非スプログ	—	低信頼度-スプログ/非スプログ	—	バランス-スプログ/非スプログ	

いては，2007〜2008 年の収集したデータセット中のスプログとの間でリンク先を共有するブログサイトを収集した(注9)．具体的には，まず，2007〜2008 年のデータセット中のスプログの HTML ファイルより，アウトリンクとなっている URL を抽出し，その中から以下の条件を満たす URL をブラックリストとして選定した．

i)　2007〜2008 年のデータセット中の非スプログの HTML ファイルのいずれにも含まれない URL である．

ii)　2007〜2008 年のデータセット中のスプログの HTML ファイルの中で，2 回以上出現する URL である．

この結果，約 5,000 件のブラックリスト URL を取得し，各ブラックリスト URL をアウトリンクとして含むブログサイトを収集した．このうち，本論文執筆時点において，人手によるスプログ/非スプログ判定を経たブログサイトの数を表 1 に示す．

3. スプログ検出のための素性

本節では SVM によるスプログ判定において用いる素性について述べる．これらの素性全てについて，スプログ/非スプログ判定における性能が評価済みであり，5. 節の評価においては，最も高い性能を示す素性の組み合わせを用いた評価結果を載せている．

(注9)：この方法では，2007〜2008 年に収集されたスプログと同種のスプログのみが収集される可能性があるが，5.2 節に示す評価結果をふまえれば，二つのデータセット中のスプログの間には十分な違いが認められると考えられる．

3.1　ブラックリスト/ホワイトリスト URL 素性

訓練事例として，スプログ/非スプログが与えられると，その HTML ファイルからアウトリンクとなっている URL を抽出する．その中から以下の条件を満たすものを選定し，ホワイトリスト URL とした．

i)　訓練事例中のスプログの HTML ファイルのいずれにも含まれない URL である．

ii)　訓練事例中の非スプログの HTML ファイルの中で，2 回以上出現する URL である．

2007〜2008 年に作成した日本語スプログ/非スプログデータセット [13], [14](2. 節) からは，約 13,000 のホワイトリスト URL が取得できた．次に，各ホワイトリスト URL u に以下のように重みづけを行い，ホワイトリスト URL 素性の値を算出した．

$$\log \sum_{u} \left(\begin{array}{c} \text{訓練事例全体の中の} \\ \text{非スプログにおける} \\ u \text{ の総出現頻度} \end{array} \right) \times \left(\begin{array}{c} \text{評価事例} \\ \text{における } u \text{ の} \\ \text{出現頻度} \end{array} \right)$$

一方，ブラックリスト URL についても，同様の手順で選定し，約 5,000 件のブラックリスト URL を取得し，スプログ判定のためにブラックリスト URL 素性の値を算出した．しかし，ブラックリスト URL 素性は，スプログ/非スプログ判定における最適な性能を達成した素性の組み合わせの中には含まれなかった．

3.2　名詞句素性

[13]〜[15] の知見より，スプログおよび非スプログ中における単語の分布には異なりがあり，特定の種類の単語は非スプロ

グよりもスプログに現れやすいということがわかっている．そこで，特定の名詞句とスプログ，非スプログとの間の相関をとらえるために，名詞句素性を導入する．

具体的には，スプログ/非スプログの本文テキストを形態素解析(注10) した結果から名詞句を抽出し，以下の分割表にしたがって，訓練データ中のスプログ/非スプログにおける名詞句 w の出現頻度を用いて，スプログと名詞句 w との間の ϕ^2 統計量を求めた．

	w	$\neg w$
訓練データ中のスプログ	$freq$(スプログ, w) $= a$	$freq$(スプログ, $\neg w$) $= b$
訓練データ中の非スプログ	$freq$(非スプログ, w) $= c$	$freq$(非スプログ, $\neg w$) $= d$

$$\phi^2(\text{スプログ},\ w) = \frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

2007~2008 年に作成された日本語スプログ/非スプログデータセットからは，約 462,894 の名詞句を収集した．また，評価事例に対しては，この名詞句素性の値として以下の式を用いた．

$$\log \sum_w \phi^2(\text{スプログ},\ w) \times \left(\begin{array}{c} \text{評価事例における} \\ w \text{ の出現頻度} \end{array} \right)$$

3.3 アンカーテキスト名詞句・リンク URL 素性

ブラックリスト/ホワイトリスト URL 素性および名詞句素性よりもより詳細な条件を設定することにより，より有効な性能を示す素性として，アンカーテキストの名詞句およびそのリンク先 URL の (緩い) 組み合わせを用いる．以下，まず，名詞句 w およびブログサイト s に対して，以下の尺度 $AncfB(w,s)$ および $AncfW(w,s)$ を定義する．

$$AncfB(w,s) = \left(\begin{array}{c} \text{ブログサイト } s \text{ 中で名詞句 } w \text{ が} \\ \text{アンカーテキストに含まれ} \\ \text{そのリンク先がブラックリスト} \\ \text{URL もしくは訓練事例中の} \\ \text{スプログとなっている回数} \end{array} \right)$$

$$AncfW(w,s) = \left(\begin{array}{c} \text{ブログサイト } s \text{ 中で名詞句 } w \text{ が} \\ \text{アンカーテキストに含まれ} \\ \text{そのリンク先がホワイトリスト} \\ \text{URL もしくは訓練事例中の} \\ \text{非スプログとなっている回数} \end{array} \right)$$

そして，訓練事例中のスプログ全体の中での総出現頻度 $\sum_s AncfB(w,s)$ が 2 以上であるものを選定し，「ブラックリスト URL へのアウトリンクを持つスプログアンカーテキスト名詞句」とする．2007~2008 年に作成した日本語スプログ/非スプログデータセットからは，約 2,000 件の「ブラックリスト URL へのアウトリンクを持つスプログアンカーテキスト名詞句」が選定された．次に，w を「ブラックリスト URL へのアウトリンクを持つスプログアンカーテキスト名詞句」として，評価事例 t に対して以下の重みを算出し，評価事例 t に対する「ブラックリスト URL へのアウトリンクを持つアンカーテキスト名詞句素性」の値とする．

$$log \sum_w \Big(\sum_{\substack{\text{訓練事例中の} \\ \text{スプログ } s}} AncfB(w,s) \Big) \times AncfB(w,t)$$

同様の手順で，訓練事例中のスプログ全体の中での総出現頻度 $\sum_s AncfW(w,s)$ が 2 以上であるものを選定し，「ホワイトリスト URL へのアウトリンクを持つスプログアンカーテキスト名詞句」とする．2007~2008 年に作成した日本語スプログ/非スプログデータセットからは，約 320 件の「ホワイトリスト URL へのアウトリンクを持つスプログアンカーテキスト名詞句」が選定された．次に，w を「ホワイトリスト URL へのアウトリンクを持つスプログアンカーテキスト名詞句」として，評価事例 t に対して以下の重みを算出し，評価事例 t に対する「ホワイトリスト URL へのアウトリンクを持つアンカーテキスト名詞句素性」の値とする．

$$\log \sum_w \Big(\sum_{\substack{\text{訓練事例中の} \\ \text{スプログ } s}} AncfW(w,s) \Big) \times AncfW(w,t)$$

3.4 リンク構造素性

アウトリンク総数，ブログサイトからのアウトリンクの中の最大リンク数，他のブログサイトとの間の相互リンク数といった，リンク構造を検出するための素性を導入した．しかし，いずれのリンク構造素性も，スプログ/非スプログ判定における最適な性能を達成した素性の組み合わせの中には含まれなかった．

4. スプログ検出および信頼度尺度

4.1 SVM を用いたスプログ検出

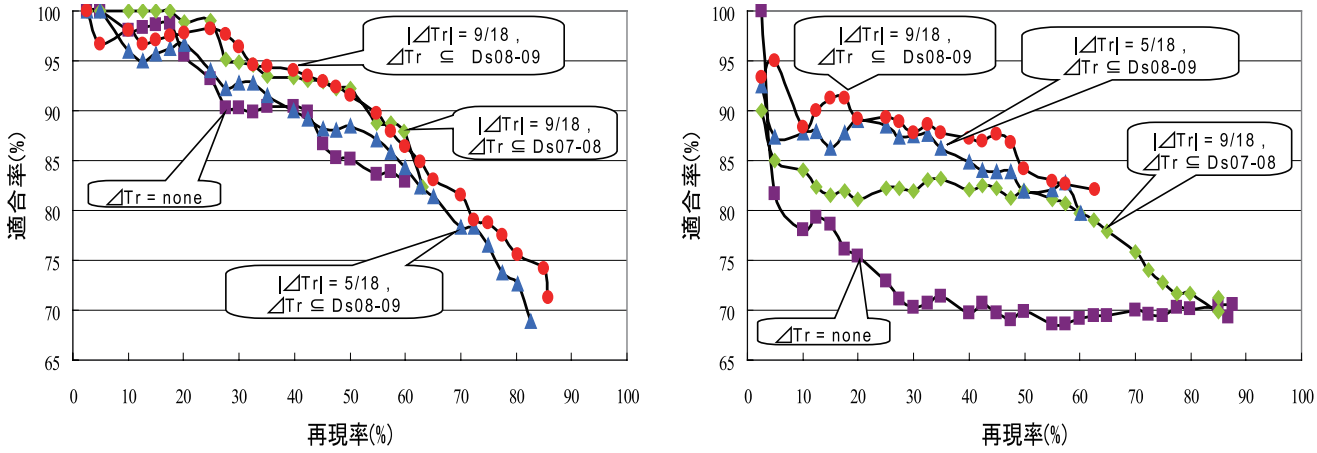
SVM 機械学習を行うためのツールとして，TinySVM (`http://chasen.org/~taku/software/TinySVM/`) を用いた．カーネル関数としては，線形および 2 次多項式を比較し，2 次多項式の方が性能が良かったため，5. 節においては，2 次多項式カーネルを用いた場合の結果を示す．また，全ての素性に値がないものは訓練データから除外する．

4.2 信頼度尺度

SVM 機械学習での信頼度尺度として，分離平面から各評価事例への距離を用いた [12](注11)．具体的には，スプログとして判定される事例に対する分離平面からの距離の下限 LBD_s，および，非スプログとして判定される事例に対する分離平面からの距離の下限 LBD_{ab} をそれぞれ設定する．

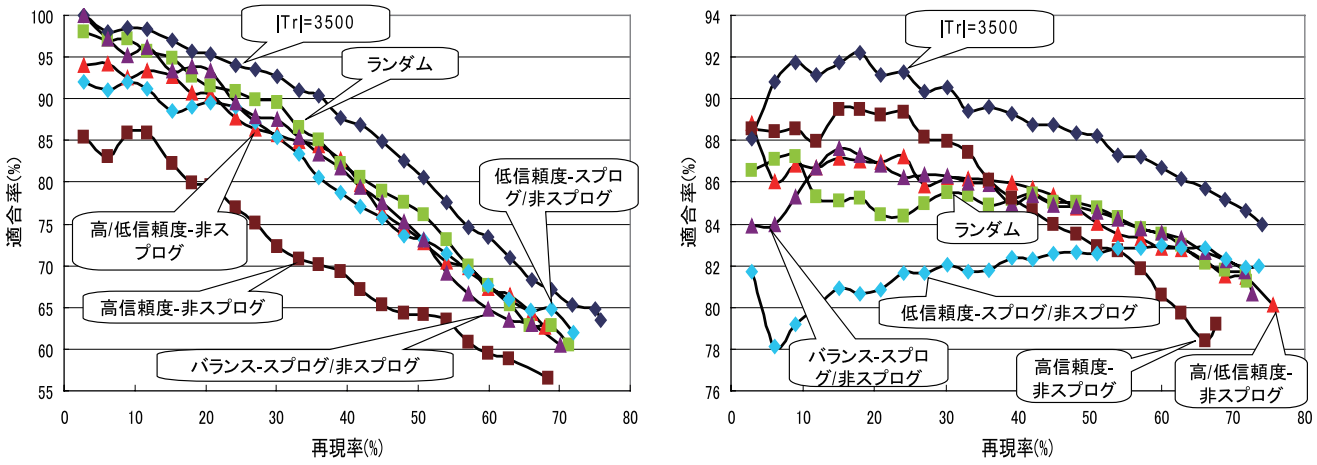
(注10)：日本語形態素解析器 茶筌 (`http://chasen-legacy.sourceforge.jp/`) および ipadic 辞書を用いた．

(注11)：機械学習および統計的自然言語処理の分野における能動学習 (例えば, [12], [16], [17] 等) 手法の研究事例においては，未知事例のうちで信頼度の低い事例を選別して訓練事例に追加する過程において，信頼度尺度が利用される．



(a) スプログ検出　　(b) 非スプログ検出

図 1　スプログ/非スプログ収集時期とスプログ検出性能の相関 (訓練データ: Ds07-08 (2007∼2008 年に作成したデータセット) の $\frac{9}{18} + \Delta Tr$, 評価データ: Ds08-09 (2008∼2009 年に作成したデータセット) の $\frac{1}{18}$)



(a) スプログ検出　　(b) 非スプログ検出

図 2　能動学習の評価：スプログ/非スプログを 1000 サイトを加えたときの再現率/適合率曲線

4.3 能動学習

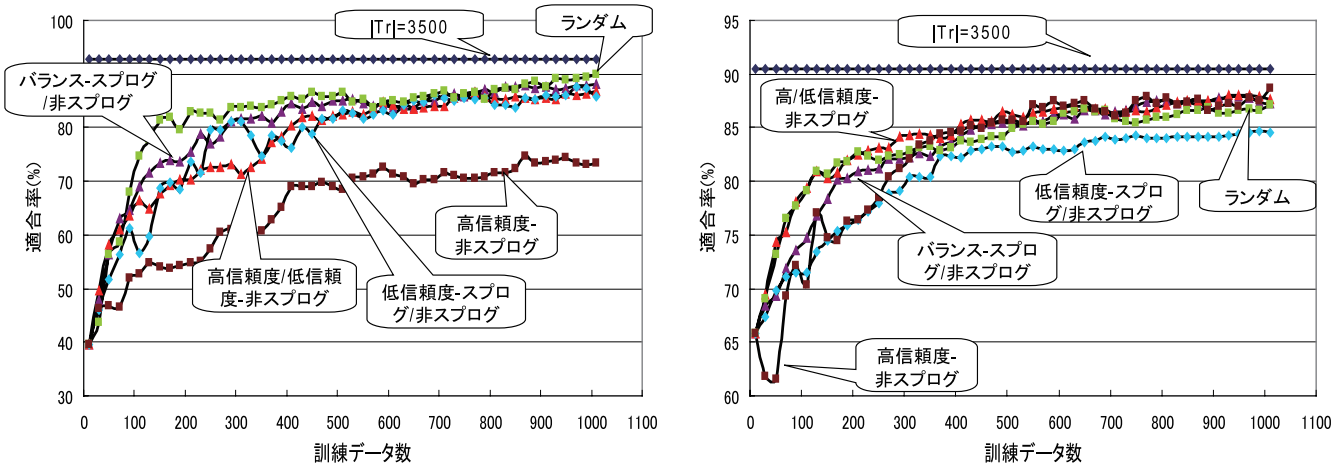
本論文では，新規のスプログ/非スプログに対して判定作業を行う際の作業量の削減を目的として，能動学習の枠組みを適用する．訓練開始時はランダムに 10 事例 (スプログ 4 事例+非スプログ 6 事例) を選び訓練データとし，390 事例を評価用データとする．表 1(c) に示す通り，ラベルなし事例 3504 事例の集合をプールとし，能動学習の各ステップににおいて，プールから 4 事例を選択的サンプリングして，人手でラベルをつけ訓練データに加える．訓練事例数が 1,010 個になるまで，この能動学習のステップを行う．

能動学習の各ステップにおいて，訓練事例に追加する事例をプールから選択する際には，分離平面から遠い距離に位置する高信頼度事例を選択するか，逆に，分離平面の近傍に位置する低信頼度事例を選択するか，の選択肢がある．(機械学習および統計的自然言語処理の分野における能動学習 (例えば, [12], [16], [17] 等) 手法の研究事例においては，低信頼度事例を選択する方式が効果的であると報告されている．) また，この選択肢とは直交する指針として，訓練事例に追加する事例をプールから選択する際に，能動学習の各ステップにおける SVM 分類器の分離平面から見て，スプログ側に位置する事例を選択するか，逆に，非スプログ側に位置する事例を選択するか，の選択肢がある．

本論文では，まず，第一の,「高信頼度事例/低信頼度事例」の選択肢については，以下の四通りの方式を評価する。

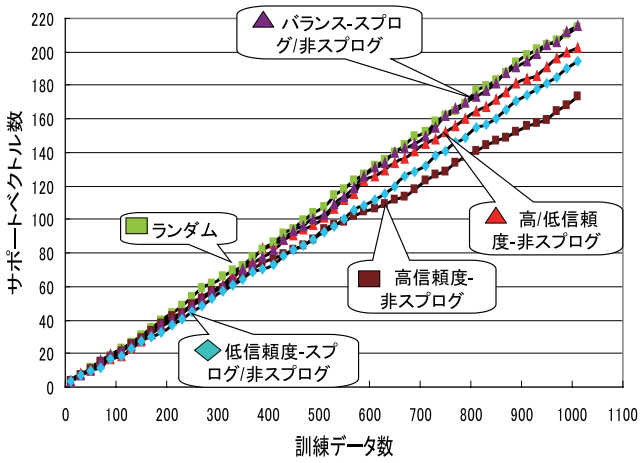
- 最も高信頼度のものを選択する (表 2 の「高信頼度」)
- 最も低信頼度の低いものを選択する (表 2 の「低信頼度」)
- 最も高信頼度のものと最も低信頼度のものを同数選択する (表 2 の「高/低信頼度」)
- 信頼度の点からみて，選択される事例が最も分散されるように事例を選択する (表 2 の「バランス」)

一方，第二の,「スプログ側/非スプログ側」の選択肢については，以下の三通りの方式を評価する。

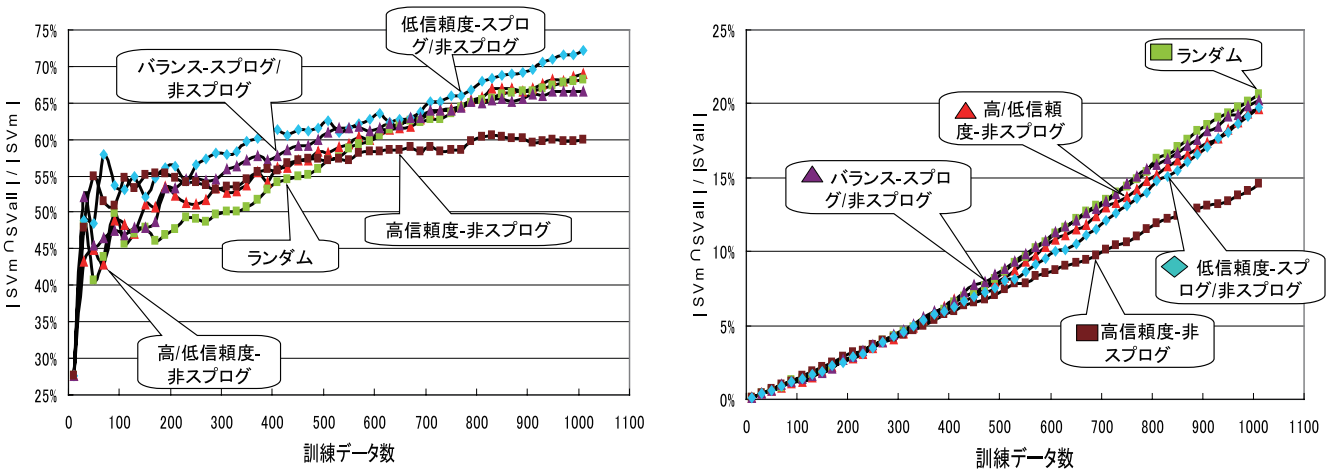


(a) スプログ検出　　(b) 非スプログ検出

図 3　能動学習の評価：再現率が 30%以上となる点における最大適合率の変化



(a) サポートベクトル数



(b) $|SVm \bigcap SVall| / |SVm|$　　(c) $|SVm \bigcap SVall| / |SVall|$

図 4　能動学習の評価：サポートベクトル数

- スプログ側から選択する (表 2 の「スプログ」)
- 非スプログ側から選択する (表 2 の「非スプログ」)
- スプログ側，非スプログ側から同数選択する (表 2 の「スプログ/非スプログ」)

本論文では，第一の選択肢の各々について，第二の選択肢を比較して，能動学習終了時の性能が最もよくなるものを選

んだところ，表 2 に記載された組合わせが得られた．そこで，これらの組み合わせ，および，能動学習の各ステップにおいて，プール中の事例をランダムに選んだ場合 (表 2 の「ランダム」)，および，全事例 (3,500 サイト) を訓練事例とした場合 (| Tr |= 3500) の比較を行う．

5. 評　　価

5.1 評 価 尺 度

本論文の評価においては，評価用スプログ事例集合，および，スプログとして判定される事例に対する分離平面からの距離の下限 LBD_s を用いて，分離平面からの距離が LBD_s 以上となる評価事例に対して，スプログとして判定した場合の再現率，適合率を測定する．そして，LBD_s を変化させた場合の再現率，適合率の推移をプロットする．同様に，評価用非スプログ事例集合，および，非スプログとして判定される事例に対する分離平面からの距離の下限 LBD_{ab} を用いて，分離平面からの距離が LBD_{ab} 以上となる評価事例に対して，非スプログとして判定した場合の再現率，適合率を測定する．そして，LBD_{ab} を変化させた場合の再現率，適合率の推移をプロットする．

5.2 スプログ/非スプログ収集時期とスプログ検出性能の相関

本節では，2007∼2008 年に収集したスプログ/非スプログデータセットと 2008∼2009 年に収集したスプログ/非スプログデータセットの 2 つのデータセットを用いて，評価用データセットよりも以前のスプログを用いて訓練した分類器，および，評価用データセットと同時期のスプログを用いて訓練した分類器の間で性能を比較することにより，一定期間の間にスプログが変化していることを示す．

まず，2007∼2008 年に作成されたデータセットのみで訓練した分類器と，2 つのデータセットを混合して訓練した分類器の 2 つの分類器を作成した(注12)．そして，2008 年 ∼2009 年に作成したデータセットから別途作成した評価データ (スプログと非スプログの割合は 1 対 1) を対象として，2 つの分類器の性能を評価した．

図 1 において，「ΔTr なし」，および，「$|\Delta Tr| = \frac{9}{18}$, $\Delta Tr \subseteqq$ Ds07-08」のプロットは，いずれも，2007∼2008 年に作成されたデータセットのみを用いて訓練した分類器の性能を示す．ここで，前者の訓練事例数は後者の半数である．一方，「$|\Delta Tr|=\frac{9}{18}$, $\Delta Tr \subseteqq$ Ds08-09」，および，「$|\Delta Tr|=\frac{5}{18}$, $\Delta Tr \subseteqq$ Ds08-09」のプロットは，時期の異なる二種類のデータセットを混合した訓練事例を用いて訓練した分類器の性能を示す．この比較からわかるように，時期の異なる二種類のデータセットを混合した訓練事例を用いて訓練した分類器の性能の方が高くなった．これは，訓練用データセット (の一部) と評価用データセットの作成時期が一致したためであると考えるのが自然である．さらに，2008∼2009 年に作成されたデータセットからの訓練事例数を多くすると，性能が向上するという結果が得られた．

以上の結果から，スプログは，時間 (今回の場合は約 1 年という時間) とともに変化していると推定され，単純に，過去のスプログを収集して SVM の訓練を行っただけでは，同数の新規のスプログを用いて訓練した分類器には及ばないことがわかった．

5.3 能動学習による選択的サンプリング方式の評価

本節では，能動学習による精度の変化と，訓練データの追加方法の比較について検証する．

まず，能動学習終了時 (新規訓練事例を 1,000 事例追加時) における再現率/適合率曲線を図 2 に示す．また，能動学習の各ステップにおいて，再現率が 30%以上となる点における最大の適合率をプロットした結果を図 3 に示す．

これらの比較からわかるように，信頼度が平均的になるように新規訓練事例を選択する「バランス」，および，「高/低信頼度」が高い性能を示した．一方，「高信頼度」や「低信頼度」の新規訓練事例選択方式の性能はよくない．この結果から，新規訓練事例への人手判定作業を削減するためには，分離平面の距離を用いて信頼度が平均的になるように事例を選択する方式が最も効果的であると考えられる．また，全事例 (3,500 サイト) を訓練事例とした場合との比較から分かるように，新規訓練事例の追加数が 1,000 個の段階では，再現率/適合率曲線は収束していないが，図 3 から分かるように，緩やかに収束する方向で推移していると言える．

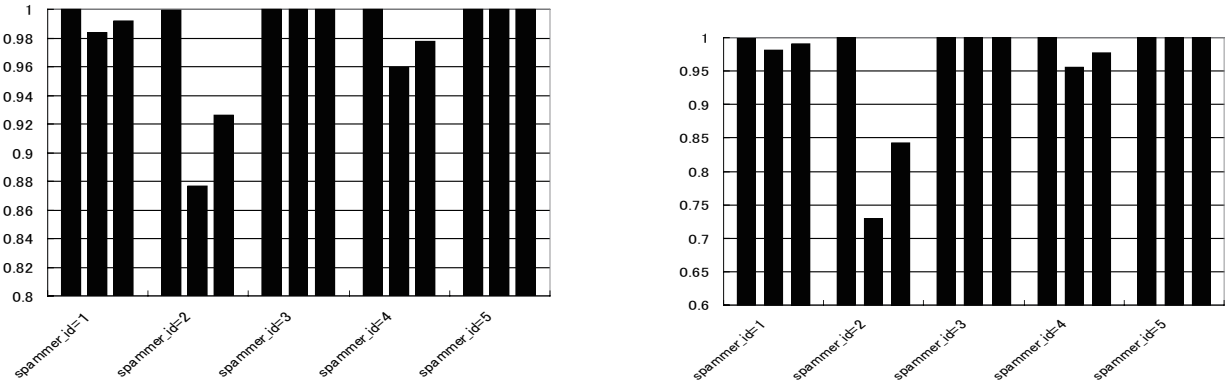
最後に，サポートベクトルの数の変化を図 4 に示す．SVM の学習においては，分離平面の位置はサポートベクトルのみによって決まっており，学習タスクの困難さの度合いを知る目安の一つとして，サポートベクトルの数が用いられる．図 4(a) の結果から分かるように，新規追加訓練事例数に対して，サポートベクトルの数は線形に増加している．また，性能のよくなかった「低信頼度」,「高信頼度」はサポートベクトルの数が少ないことがわかり，サポートベクトルの数と判定性能の間に相関があることが分かる．また，図 4(b)，(c) より，$|SVm \bigcap SVall| / |SVm|$ はやや収束気味であるが，$|SVm \bigcap SVall| / |SVall|$ は線形に増加し続けている．

以上の結果は，新規訓練事例の追加数が 1,000 個の段階で，再現率/適合率曲線が緩やかに収束する方向で推移していることと比べて対照的であると言える．つまり，学習結果である分離平面を表現するにあたっては，できる限り多数のサポートベクトルを使用するが，その追加数に合わせて性能が線形に改善するわけではない．

5.4 大量生成型スプログに対する評価

[13], [14] においては，同一のスパマーによって作成されたと推定される複数のスプログを「大量生成型」スプログ，また，その作成者を「大量生成型」スパマーと定義し，データセット中のスプログに対して，大量生成型スパマーの ID を付与している．そこで，次に，大量生成型スパマーの ID 別に，各スパマーが識別可能かどうかの評価実験を行った．ここでは，データセット中に十分な数含まれていた ID=1∼5 の大量生成型ス

(注12)：訓練および評価に用いたデータの数を，表 1 (b) に示す．表 1 (a) の全データのうち，2007∼2008 年に作成されたデータセットと 2008∼2009 年に作成されたデータセットの双方から，同サイズの 2 つの訓練用データセットを作成する．ここで，この 2 つの訓練用データセットにおいて，スプログと非スプログの割合を 1 対 1 とする．



(a) 全タイプの大量生成型スプログを用いて訓練　　(b) 各タイプの大量生成型スプログを除いて訓練

図 5　大量生成型スプログの検出性能 (左から適合率，再現率，F 値)

パマーを対象とした．全タイプの大量生成型スプログを用いて訓練し，各タイプの大量生成型スプログの識別を評価した場合を図 5 (a) に，また，各タイプの大量生成型スプログを除いて訓練し，同一タイプの大量生成型スプログの識別を評価した場合を図 5 (b) に，それぞれ示す．

他のタイプに比べると，ID=2 のタイプの大量生成型スプログだけは，識別性能が低く，しかも，同一タイプの大量生成型スプログを除いた場合にも，他のタイプと比較して，識別性能の低下幅が大きかった．

今後は，大量生成型スプログの各タイプの同定について評価を行う．

6. おわりに

本論文では，機械学習のひとつである SVM を用いた枠組みによって，スパムブログ (スプログ) の判定を行うタスクを設定した．未判定のブログが入力されたとき，SVM の分離平面によって，スパムブログかそうでないかを決定する．さらに，SVM では分離平面との距離を出力できるので，分離平面との距離を信頼度として利用する方式について述べた．また，前後する 2 つの期間のデータセットを用いて，評価用データセットよりも以前のスプログを用いて訓練した分類器，および，評価用データセットと同時期のスプログを用いて訓練した分類器の間で性能を比較することにより，一定期間の間にスプログが変化していることを示した．また，能動学習の枠組みにより，訓練データの追加の仕方として，信頼度が平均的になるように追加していく手法の有用性を示した．今後は，先行研究 [4]〜[6] に挙げられるように，ブログ記事の投稿頻度などを素性として追加することにより，性能を改善する．

文　献

[1] T. Nanno, T. Fujiki, Y. Suzuki, and M. Okumura. Automatically collecting, monitoring, and mining Japanese weblogs. In *WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pp. 320–321. ACM Press, 2004.

[2] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *AIRWeb '05: Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web*, pp. 39–47, 2005.

[3] *Wikipedia, Spam blog.* `http://en.wikipedia.org/wiki/Spam_blog`.

[4] P. Kolari, A. Joshi, and T. Finin. Characterizing the splogosphere. In *Proc. 3rd Ann. Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.

[5] C. Macdonald and I. Ounis. The TREC Blogs06 collection : Creating and analysing a blog test collection. Technical Report TR-2006-224, University of Glasgow, Department of Computing Science, 2006.

[6] P. Kolari, T. Finin, and A. Joshi. Spam in blogs and social media. In *Tutorial at ICWSM*, 2007.

[7] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. L. Tseng. Splog detection using self-similarity analysis on blog temporal dynamics. In *AIRWeb '07: Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*, pp. 1–8, 2007.

[8] 石田和成．スパムブログの推定と抽出．日本データベース学会 Letters, Vol. 6, No. 4, pp. 37–40, 2008.

[9] 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．Web とデータベースに関するフォーラム (WebDB Forum)2008 論文集．情報処理学会, 2008.

[10] P. Kolari, T. Finin, and A. Joshi. 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.

[11] V. N. Vapnik. *Statistical Learning Theory.* Wiley-Interscience, 1998.

[12] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *Proc. 17th ICML*, pp. 999–1006, 2000.

[13] 佐藤有記, 宇津呂武仁, 福原知宏, 河田容英, 村上嘉陽, 中川裕志, 神門典子．キーワードの時系列特性を利用したスパムブログの収集・類型化・データセット作成．DEWS2008 論文集, 2008.

[14] Y. Sato, T. Utsuro, T. Fukuhara, Y. Kawada, Y. Murakami, H. Nakagawa, and N. Kando. Analysing features of Japanese splogs and characteristics of keywords. In *Proc. 4th AIRWeb*, 2008.

[15] Y.M. Wang, M. Ma, Y. Niu, and H. Chen. Spam double-funnel: Connecting web spammers with advertisers,. In *Proc. 16th WWW Conf.*, pp. 291–300, 2007.

[16] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proc. 17th SIGIR*, pp. 3–12, 1994.

[17] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *Proc. 17th ICML*, pp. 839–846, 2000.