# スプログ監視支援のための信頼度つきスプログ検出ツールの開発

## 1. 背景と目的

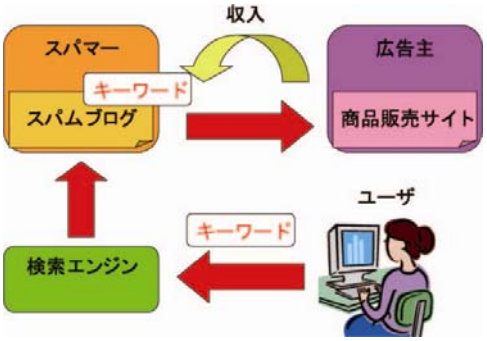


図 1　スプログのシステム

Web上の情報が質・量ともに充実すると同時に、その情報の検索技術も向上し、Web空間の情報が生活にも密接なものとなっていくにつれ、CGM(Consumer Generated media)のマーケティングにおける利用価値が高まり、重要性が認められてきている。しかし、図1で示すようなスプログに代表される粗悪な情報の存在が、特定対象への不当な利益・不利益を招いたり、本当に必要な情報へ到達する上で障害となる。Niftyの調査によると、ブログ空間中のスプログの混入率は 4 割にのぼる。(http://it.nikkei.co.jp/internet/news/index.aspx?n=MMITba000009042008)

スプログ混入の弊害はマーケティング調査に如実に表れる。例えば、図2の「SOY JOY」に関するブログの更新頻度を、仮にそのまま一般の関心と読み替えた場合、スプログのみでバーストを起こしている部分を関心のバーストと読み違えてしまうと、大きな誤認識となってしまう。

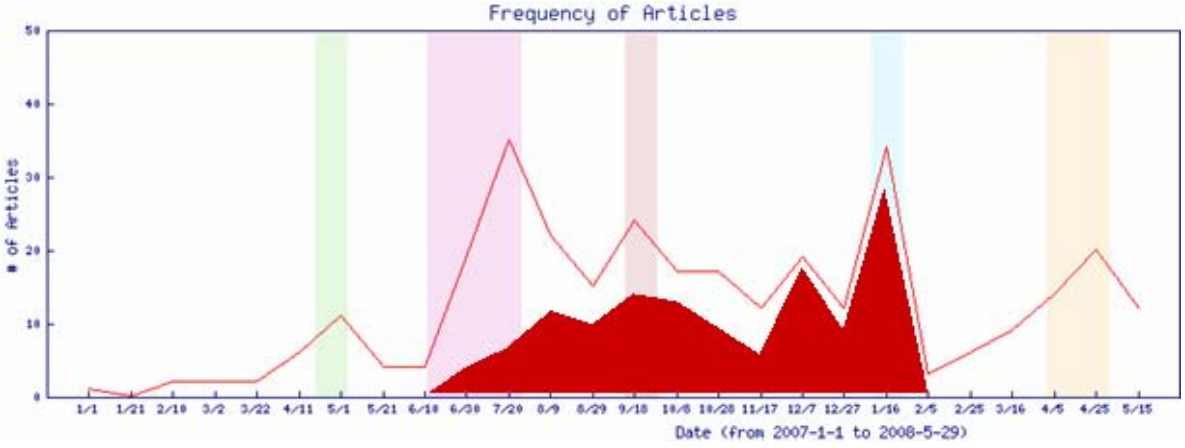


図 2 「SOY JOY」キーワードの更新頻度推移グラフ(赤ベタ部はスプログ分)

このような情報をフィルタリングするために、図3に示す人手による監視を行うサービスが現われ始めているが、Web空間の大規模な領域を監視するのは困難である。信頼性を保ちながら、なるべく大きな領域を監視するというエンジンが、こうしたサービスにおいて大きな価値を持つようになるだろう。

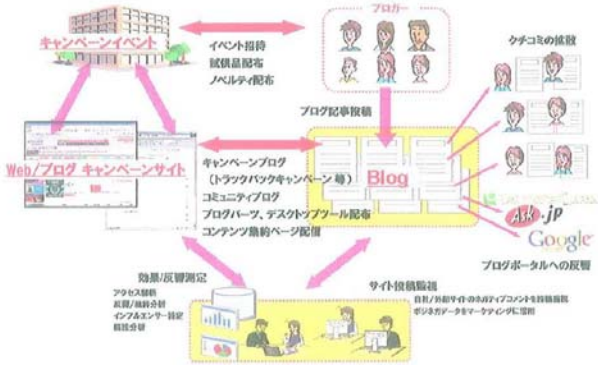


**図 3 スプログ監視ビジネスモデル**

スプログ監視サービスにおいて特定のCGMがスプログであるかどうかを人手で判定する作業は膨大な手間と時間がかかる。この問題を解決するために対象CGMの質を信頼度つきで提示するシステムを開発することで、判定信頼度の低いグレーゾーンを明確にし、監視対象を絞ってスプログ監視を効率化する。

本プロジェクトで開発するスプログ監視システムは監視対象のブログが入力されるとスプログ/非スプログの判定結果を、信頼度を付加して出力する。システムの出力結果が一定の閾値以上か以下かによって監視対象のブログを高信頼度/低信頼度のグループに選別する。高信頼度のグループは精度が 95%以上となるようにして、作業者が監視をする必要がないようにする。低信頼度のグループのみを作業者が監視することで、システム運用の作業効率を高める。低信頼度への出力を 25%以下に抑えることを開発目標とする。本システムはビジネス展開を主目的としており、CGM監視サービスに実用的なシステムを開発することを想定している。

## 2. 開発の内容

**動作環境**

Apache2.2 **以上**

Tomcat5.5 **以上**

MYSQL

**本システムの開発機能を以下に記す。**

### 2.1 機械学習によるスプログ判定器の生成

**機械学習にはオープンソースのサポートベクターマシン TinySVM を用いた。**

SVM における分離平面からの距離を、各判定での「信頼度」として（図 4)、これを判定とともに出力した。

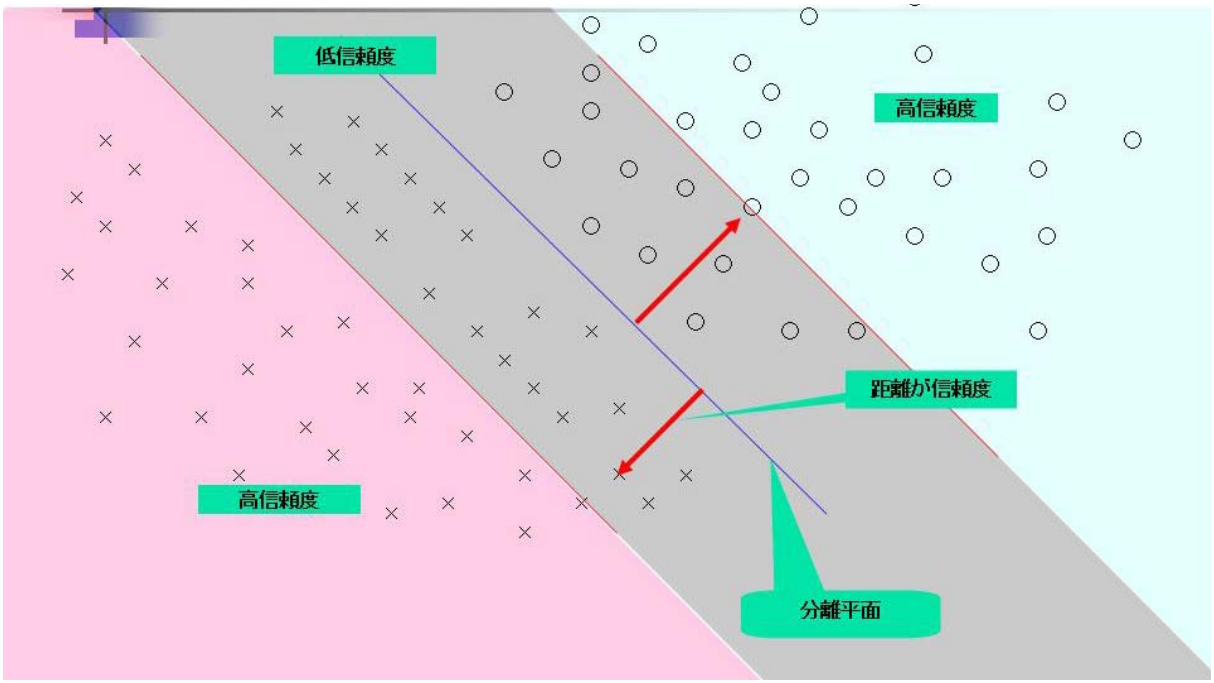


図 4　SVM における分離平面と信頼度尺度

「信頼度」をランキングし、上位 90%の部分を「高信頼度」、下位 10%を「低信頼度」として判定対象ブログを 2 つのグループに分け、低信頼度部分だけを人手で判定する。

### 2.2 インタフェース

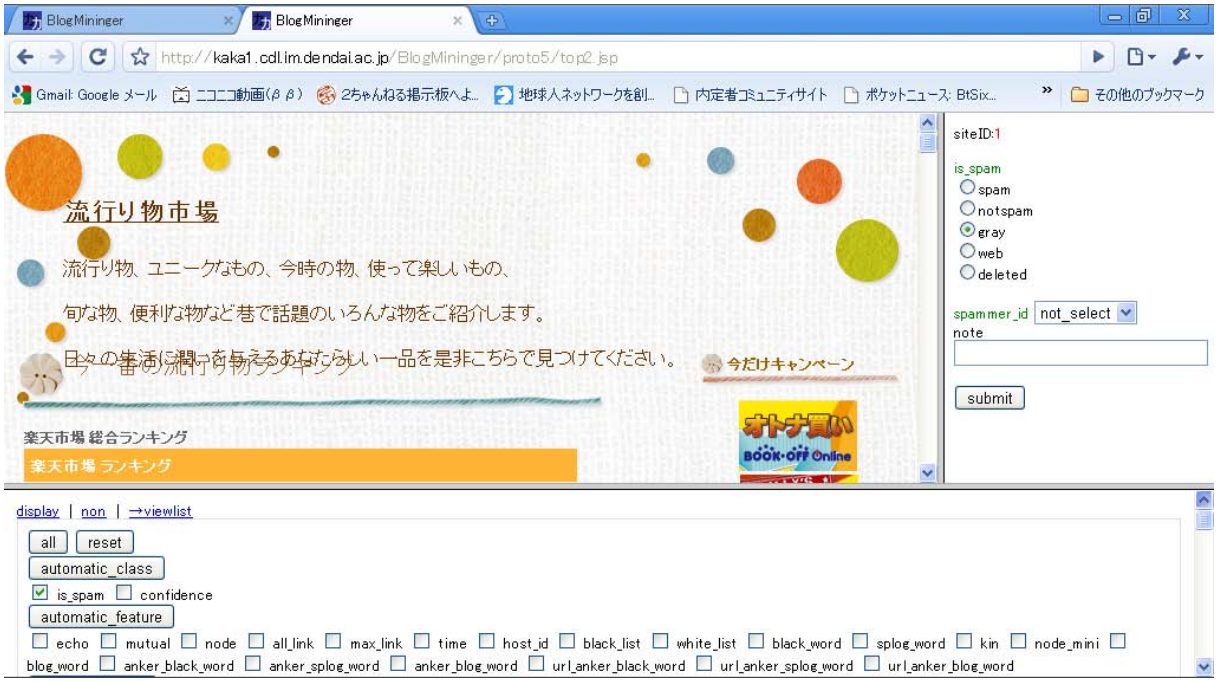


図 5　システムインタフェースのスクリーンショット


図 6　作業者の判定結果入力フォーム

| site_id | file_id | code | timestamp | keyword | keyword_id | fid |
|---|---|---|---|---|---|---|
| 1 | 1151547 | http://blog.livedoor.jp/yamo777 | 2007-09-28 13:34:49.0 | ダイエット | 23 | 28441 |

| is_spam | confidence |
|---|---|
| 1 | 954978 |

| site_id | file_id | code | timestamp | keyword | keyword_id | fid |
|---|---|---|---|---|---|---|
| 2 | 948726 | http://d.hatena.ne.jp/sadn | 2007-10-09 09:55:06.0 | youtube | 31 | 38313 |

| is_spam | confidence |
|---|---|
| 0 | -1457560 |

| site_id | file_id | code | timestamp | keyword | keyword_id | fid |
|---|---|---|---|---|---|---|
| 3 | 415226 | http://ssll.nablog.net/blog | 2007-10-01 18:27:06.0 | 健康食品 | 21 | 55782 |

| is_spam | confidence |
|---|---|
| 1 | 1073180 |

**図 7　機械判定の結果**

スプログデータベースの収集作業を、実際に作業者が行うにあたって、データベースにアクセスするためのインタフェースを作成した。図 5 に本システムのスクリーンショットを示す。

図 5 の下部フレームにおいて、データベース中の未判定のブログデータの中から、一定の特徴を持つものを検索し、上左部にブログの実体を表示する。作業者は、

- **スプログである/スプログでない/判別できないのいずれに該当するかを判定**
- **スプログであるなら、同一スパマーが作成したスプログがある場合、大量生成型として ID を付与**

という二つの判定作業の結果を、上右部の入力フォーム（図 6）にて選択・入力できる。

本システムは図 7 のように機械判定の結果と信頼度を検索、出力できるようになっている。実際の作業時には作業者が機械学習の信頼度（confidence）が一定の閾値以下であるスプログのみを表示して判定を行うことができる。

### 3. 従来の技術（または機能）との相違

1. 従来のスプログ除去においては株式会社ユーテラスなどが行っているが全て自動判定のものである。スプログは日々変化していくものであり、自動判定ツールではいずれは洩れや抜けが出てくるため人手で判定するということが重要になってくる。今現在日々変化するスプログに対応するためにスプログの監視を支援するためのツールは存在しない。また本システムは人手判定をするため、機械学習による分離平面からの距離を信頼度としその信頼度を用いて人手作業量の軽減と判定精度の向上を可能にした。
2. 同一スパマーが作成したスプログが共有する素性の検出を実現。さらに、人手判定において同一スパマーが作成したスプログに ID を付与し、作業者間で情報を共有することにより、同一スパマーが作成したスプログを網羅的に判定することが容易になり、人手での誤判定の減少を可能になった。また、以前は作業管理者が人手で同一スパマーが作成したスプログのタイプをチェックしながら集計をしていた作業がなくなった。

### 4. 期待される効果

このシステムの利用目的として、2 通り想定する。

- ツールとして、ブログホスト会社や検索エンジン会社に使用してもらう
- 背景と目的に記したようなスプログ監視ビジネスの展開に利用する

ビジネス展開として 2 通りの戦略が考えられる。

- 大規模なブログ空間を監視するサービス

自ら大規模なブログ空間からデータを収集し、スプログを判定する

- 決められたブログ空間を監視するサービス

ブログ会社などが既に持っているデータを使用し、スプログを判定する
どちらのビジネス展開の戦略をとるにしても、大規模なデータを扱うことが必須になってくる。一方、計算資源は有限であるため、どのようにブログを選ぶのかが重要になってくる。今後大規模なブログデータからどのような手法でブログを選択すれば異なったスプログを多く選ぶことができるのかについて検討していく。
ツールとしてブログホスト会社や検索エンジン会社に使ってもらうことや、背景と目的に記したようなスプログ監視ビジネスの展開に利用することを進めることにより、次のような効果が期待できる。

- ブログホスト会社での自社のスプログの減少
- 検索会社での検索結果の質の向上
- マーケティング会社でのブログを用いた世の中の動向調査の質の向上

### 5.普及（又は）活用の見通し

ツールとしての利用者はブログホスト会社、検索エンジン会社などが考えられる。
同一スパマーが作成したスプログの ID や人手でつけたスプログの特徴などを検索、表示できる機能により、同一スパマーが作成したスプログのさまざまな情報を同時に見ることが可

能になり、どのようなスパマーか知ることができる。これにより、一つのスプログを見ただけではスプログと判定できないものの判定が可能になり、人手での誤判定が減少すると考えられる。
実際に大規模なデータを持っているブログホスト会社や検索エンジン会社などがこのシステムを使用した場合、機械学習の高信頼度グループの精度を 95%以上とすると人手判定は低信頼度部分だけでよいので作業量の軽減が可能であると考えられる。
これらのことより、ブログホスト会社や検索エンジン会社で活用されていくと考えている。

## 6. 開発者名（所属）

片山太一（筑波大学大学院システム情報工学研究科知能機能システム専攻）