IPA

2008 年度上期未踏 IT 人材発掘・育成事業(未踏ユース)採択案件評価書

1. 担当PM

筧　捷彦 PM(早稲田大学　基幹理工学部　情報理工学科　教授)

2. 採択者氏名

チーフクリエータ:　片山　太一(筑波大学第三学群工学システム学類 4 年)
コクリエータ:　なし

3. プロジェクト管理組織

株式会社 ゴーガ

4. 委託金支払額

2,991,439 円

5. テーマ名

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

6. 関連Webサイト

なし

7. テーマ概要

スプログ監視サービスに対して、CGM の人手判定をする上で、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、監視対象を絞って効率化するシステムを作る。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。自社ブログへの監視ビジネスは既に存在しており、運営会社規模へのビジネスを見込んでいるナビックスとの連携は既に確立している。

より具体的には、ナビックスで作業者を雇用して、このシステムを扱わせるような運用の仕方を想定し、監視対象のブログを入力することで、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をする。この出力を信頼度において一定の閾値で高信頼度/低信頼度のグループにわける。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。

開発手法の概要は、現在所属している筑波大学システム情報工学研究科の研究資源を受け継ぐ形で、日本語スプログデータセットとその分析プログラムを用いて、このデータセットをデータベース管理し外部公開しつつ、他の研究環境においても独自にデータの拡張作業ができるようなプラットフォームを完成させることである。

8. 採択理由

Web 空間が質・量とも膨大なものになり、しかも検索技術も向上したことから、CGM(Consumer Generated Media)が商業利用の対象となり、ブログサイトの中にも、高い検索順位を得ることだけを目的とした"スパム"ブログ（スプログ）が多量に生まれてきている。これらのスプログをフィルタリングしてしまうための仕組みづくりには、高精度のスプログ検出ツールが不可欠である。このプロジェクトは、信頼度つきのスプログ判定ツールを開発して、スプログ監視ビジネスへの展開を図るものである。
スプログ判定を行うのに使われている手法は、人手によってスプログと判定したデータを集めて、そのデータセットを対象として特徴抽出を行い、学習させてフィルタを構成するというものである。こうしたフィルタの精度を95%にまであげたい、というのがこのプロジェクトの目標である。
提案者は既に、人手で 110url×50 キーワード＝5500url ほどのスプログ・非スプログと判定されたデータを収集している。
そのデータをさまざまに調べてみて、スプログの周りのリンクの構造が、通常のブログのリンクとは際立った違いをもっていることを発見している。
このように、スプログと判定したものを集めたデータセットを使ってスプログを自動収集し、集まったデータそれぞれにそのスプログ「精度」を何らかの形で評価する。その精度が低いものを対象にして、人手をかけて判定をやり直す。こうして精度をあげるとともに、精度が上がったデータセットに対して、再び解析を行い、学習を行わせて、つぎのデータ収集にあたる。こうしたプロセスを継続的に行っていく。
得られているデータセットは、さらに様々に解析して、従来のスプログ素性に近い特徴

を持つ新傾向のものや、新パターンのものを発見できるようにしていこうという計画である。きわめて実験的要素の高いものではあるが、これなしに、スプログ検出の精度をあげることはほとんど不可能でもある。未踏ユースの期間中にどれだけの発見ができるかが勝負である。開発者のこれまでの経験と知識のありったけをぶつけて成果を生み出してくれることを期待している。

9. 開発目標

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。

この目標を達成するのに，具体的にはつぎのものを開発する。

- スプログデータベースの作成
- 機械学習によるスプログ判定器の生成
- スプログからの素性抽出プログラム

これらを使うためのユーザインタフェース

10. 進捗概要

このプロジェクトは，スプログの監視がビジネスとして成立するに至っている状況を踏まえ，かつ，その監視を全自動化を図っても100 %に近い精度を上げることはできそうにないことを踏まえて，人手による判別を有効に支援するシステムを開発することを目指した。その人手による判別，という絶対的な判定結果がそこに生まれることを利用して，作成する支援システムの自動判別部の訓練にその絶対的な判定結果を使う，という基本方針をもったプロジェクトであった。

そこで，現実のスプログ監視にあたっている業者に委託して，そこで判定対象となったブログ群に人手による判別結果を添えたデータを集め，それを使って自動判別部に用いる判別データとして何を対象とするのがよいか，を実験的に分析するための実験データとして使う計画になっていた。

委託先の業者との契約が成立し，そこからデータが提供されるまでに時間がかかってしまったために，実験的に分析し，それをもとに判別部を設計し実装するのが大幅

に遅れてしまった。

いったんデータが提供されてからは，素早くさまざまな作業を進め，最終的には所期のシステムの開発とそれを使っての人手による監視作業の時間軽減の確認までを終えることができた。しかしながら，開発されたシステムのユーザインタフェースの洗練に使う時間がなくなってしまった。

11. 成果

つぎの図に示すとおりのユーザインタフェースをもったシステムとして仕上げた。

サーチエンジンを使って対象とするブログを直接に画面上で見ることのできる画面部分と，そのブログを人手でスプログか否かの判定を行った結果，およびその判定に関連するいくつもの属性を登録するための画面部分とが配置された画面構成となっている。

対象ブログを選ぶと，システムはそのブログがスプログであるかどうかを自動判定した結果を示してくれる。自動判定の結果は，“スプログ”，“非スプログ”，“判定できず”のいずれかである。この自動判定結果も参考にして，利用者は自ら判定を下す。“スプログ”，“非スプログ”の自動判定は高信頼度で得られたものである。利用者はその判定をそのまま受け入れてもいいし，自ら判定してもいい。“判定できず”のときには自ら判定しなければならない。このインタフェースを介して行った人手による最終判定結果は，自動的にスプログデータベースに登録される。さらに，スプログだと判定したときには，さらに，同一構造のスプログが他にもあるかどうかをしらべ，ある場合には“大量生成型”として ID を振るようになっている。

こうした人手による判定結果のついたブログのデータベースの内容による機械学習によって，機械的に判定を行うエンジンを開発した。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，機械学習を行わせた実験によって最高精度を示したつぎの素性を入力に用いることにして，スプログ判定器を作成した。

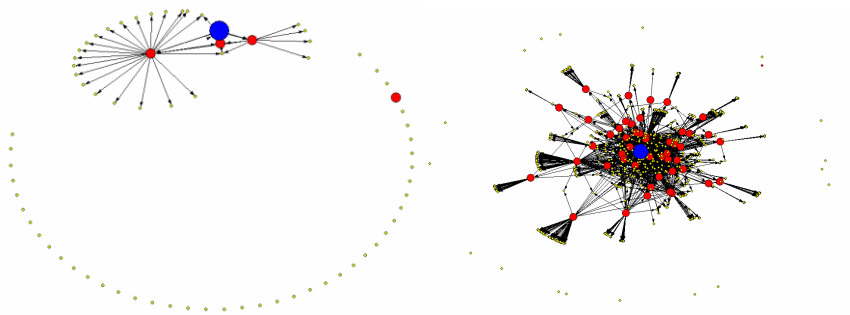
- 文字列素性
- リンク素性

このエンジンは，機械判定の結果を返すだけでなく，その判定対象となったブログがSVM における分離平面からどれだけの距離をおいた位置にあったかを，その判定の“信頼度”として返してくる。機械判定の結果すべてをこの信頼度でランク付けし，上位 90%にはいった判定結果を“高信頼度”，下位 10%のものを“低信頼度”として二分した。

高信頼度グループのブログにおいて，スプログと機械判定されたもののうちスプログであったものは 94%，非スプログと機械判定されたもののうち非スプログであったものは 96%であった。これに対して，低信頼度グループのブログのうち 74%が現実にはスプログであった。つまり，高信頼度グループに属するブログは，その機械判定の精度がほぼ95%あることになる。

機械判定に用いた文字列素性とリンク素性は，つぎのとおりである。

まず，スプログと非スプログとでは，そのリンク構造がつぎのように広がっていることが知られている。



スプログのリンク　　非スプログのリンク

すなわち，非スプログは一般ブログなどの相互リンクがあるためコミュニティを形成しやすいが，スプログはアフィリエイトサイトなどが多いためコミュニティを形成しにくいという特徴がある。これらの特徴を表すものとして，相互リンクノード数（観測対象から相互リンクのみで繋がるノード数），直接相互リンクノード数（観測対象と直接に相互リンクで繋がるノード数），リンク先ノード数（観測対象からのリンク先であるノード数），

最大リンク数（観測対象からの1 つのページへの最大リンク数），ブラックリストURL（スプログに特有の，特定のリンク先の有無），および，ホワイトリストURL（非スプログに特有の，特定のリンク先の有無）を判定データに用いた。

また，スプログには，スプログ特有の文字列が現れるという特徴がある。この特徴を文字列素性とよび，ブラックリスト名詞句（スプログに特有の，特定の形態素の出現頻度。株価，ギャンブル，ショッピングなどの用語が代表的である）および，ホワイトリスト名詞句（非スプログに特有の，特定の形態素の出現頻度。流行のトピックが多い）を当てた。ブラックリストに入れる語，ホワイトリストに入れる語は，それぞれデータベースに蓄積されたブログと非スプログについて，そこに出現する単語とその頻度を調査して定めた。

こうして出来上がったシステムを，データ収集を委託した業者に使用してもらい，その作業効率の変化を調査したところ，つぎのような結果を得た。まず，従来のエクセル表とウェブブラウザを使った作業では，１５分あたり 18.4 ブログが処理できていたのに対して，データベースと自動判定の支援を備えた開発したシステムを使った作業では，１５分あたり 28.7 ブログが処理できるようになったという。

１２．プロジェクト評価

開発を予定した機能をもったシステムが出来上がり，そのシステムを実際に使ってもらって効率の向上がもたらされたことの確認もできた，ということでは，プロジェクトは一応成功裏におわったことになる。しかしながら，開発者の秀でた部分が伸ばされたか，発揮できたか，という点では満足のいくものではなかった。

一つには，スプログの人手による判定結果のついたデータを得る作業を業者に委託するのに手間取ったため，有効な開発期間が大幅に制限されてしまったことがある。スプログの作者たちは，監視されスプログが排除されたとなると，またその監視の裏を欠く策を講じてくる。そうした長期にわたる変化にこのシステムがどのように対応できるのか，といった観測をし，手を打つことを行う時間はなかった。また，人手による判定が不可欠で，そのための支援を行うのだ，という方針からすると，ユーザインタフェースの洗練は不可欠のものであるが，それを行う時間もとれなかった。残念である。

最終の結果から見ると，開発したシステムによる単位時間あたりの処理件数は，在来のものに比べて，大幅に伸びたものの，２倍まではいかなかった。スプログ監視，ということに関しての作業効率改善を目標とするからには，その作業のボトルネックがどこにあるか，ということについての調査・分析をさらに重ねて，その結果に応じた改善方法を考える必要がありそうである。

13. 今後の課題

まず，このシステムの本格的な実用化のためには，判定アルゴリズムやインタフェースを改善する必要がある。また，スプログ監視作業について，その状況をさらに調査・分析してみる必要がありそうである。

このプロジェクトによって基本的な仕組みができた。そのデータベース中のスプログ判定作業結果をもとにして，スプログが多く含まれるブログ領域を特定することができる。開発者は，その領域のブログ記事を収集し，これまで機械判定のためのデータが不足していた大量生成型スパマーのスプログデータを収集しているところだという。これにより，大量生成型スパマーであることの判定に関しても，本システムのスプログ判定と同様に，機械的な補助をすることができるシステムとして開発を進めていく予定であるという。さらなる工夫・前進を期待したい。