ウェブサイト間の類似度を用いたウェブスパムの検出の研究

青野雅樹 (第４工学系, セマンティックアーカイブコア)

1　はじめに

われわれは、MSRC の中のセマンティックアーカイブコアの属し、主としてメディア（テキスト、画像、三次元モデル形状など）の概念的な類似度、H19 年度は主として Web 上のテキストデータに関して、集団が「タグ」として与えたソーシャルブックマークを使った関連 Web ページ検索を行ったが、H20 年度では、ウェブサイト間の類似度を用いたウェブスパムの検出を行ったのでそれを報告する。

様々なニュースサイトやブログの記事を流用することで大量のウェブページを自動的に生成するスパムは検出の難しいウェブスパムの 1 つである。そのようなスパムを検出するために、我々はウェブスパムが生成されるメカニズムに注目し、ウェブページ間の構造の類似度を求めることでウェブスパムを検出する方法を提案する。WEBSPAM-UK2007 データセット[1]を用いて実験を行った結果、提案手法を加えることでウェブスパム検出の精度が大きく向上することを確認した。また、ウェブスパム検出のワークショップである AIRWEB2008[2]の結果と比較したところ、AUC と F 値において最も良好な数値が得られた。

2　序論

2.1　ウェブスパムの定義

Gyöngyi らはウェブスパムの一般的な定義として、「ウェブページのランキングを不当に高める全ての意図的な行為」を提示し、ウェブスパムを Term Spam と Link Spam の 2 つに大別した。以下にウェブスパムの種類について説明する。

Term Spam　TFIDF は検索エンジンがウェブページの順位を決めるために使う重要な指標の 1 つがである。スパミングの観点から TFIDF スコアを考えると、IDF はスパマーにより制御できないスコアであるため、スパミングは TF を高めることで達成される。例えば、ウェブページの内容とは関係のないキーワードを追加することで、特定のクエリに対するウェブページの関連性を高める方法がある。Ntoulas らは、ウェブページの圧縮率や特定の検索クエリに対する適合率等を用いる方法が提案している [?]。

Link Spam　検索エンジンはウェブページの順位を決めるとき、TFIDF に加えリンク情報も重視する [?]。検索エンジンがリンクベースの重要度を決めるときに用いる指標の 1 つが PageRank である [?]。PageRank はウェブページが外部から得るリンクの数はそのウェブページの重要度と関連性があると仮定し、ウェブページの重要度を決定する。これは、重要度の高い複数のウェブページからハイパーリンクを得ているウェブページの重要度は高いという考えに基づく。リンクベースのスパムは、この PageRank を高めることで達成される。Link Spam の一例として Link Farm が挙げられる。Gyöngyi らは、PageRank が最適化されたグループを識別することでウェブスパムを検出する方法を提案している [?]。

3　提案手法

3.1　前処理

各類似度を求めるためには HTML ファイルに対する適切な前処理が必要となる。HTML ファイルとは HTML により記述されたファイルのことであり、ウェブブラウザによりレンダリングされる前の状態である。前処理では、この HTML ファイルを構文解析し、類似度計算に必要な情報のみを抽出することで後の処理を効率的に行う。実際には以下に示す 3 種類のドキュメントを HTML ファイルから作成する。

- *words* ドキュメント（テキスト類似度、比較用）
- *tags* ドキュメント（タグ類似度）
- *scripts* ドキュメント（スクリプト類似度）

以後、これらのドキュメントは独立に扱い、*words* ドキュメントからはテキスト類似度、*tags* ドキュメントからはタグ類似度、*scripts* ドキュメントからはスクリプト類似度を計算し、*words*、*tags*、*scripts* の 3 つの素性を構築する。以下、この 3 種類の前処理についてそれぞれ説明を行う。

words　*words* ドキュメントは HTML ファイルの HTML タグ、Javascript、コメントを除いたものである。多くのケースでウェブページの本文に相当する部分と言い替えることができる。*words* ドキュメントを抽出することで、ウェブページ間の類似度を本文に基づいて計算することが可能となる。

[1] UK2007, http://barcelona.research.yahoo.net/webspam/
[2] AIRWEB2008, http://airweb.cse.lehigh.edu/2008/

tags　*tags* ドキュメントは HTML ファイルの HTML タグと HTML タグの属性のみを抽出するものである。但し、<body> より前に出現するタグはウェブページ間での差異が小さく、差別化要因にならないため処理の対象から除外する。

scripts　*scripts* ドキュメントは HTML ファイルの <script> と <noscript> のみを抽出するものである。

3.2　類似度尺度

例として、*words* 処理により得られたドキュメント集合 D を考える。このドキュメント集合 D を n-grams によりフレーズ単位で抽出し、ユニークなフレーズ（以下、要素）に要素番号を付与する。この操作により、ドキュメント集合 D は要素番号を用いた集合 $S_D \in \{1, ..., n\}$ により表現することができる。また、n-grams を用いることで、単語の位置関係を部分的に保持した状態で要素集合を構築することが可能となる。n はドキュメント集合 D に含まれるユニークな要素数に等しい。このとき、任意のドキュメント A、B 間の類似度 $sim(A, B)$ は以下の式により定義される。

$$sim(A, B) = \frac{|S_A \cap S_B|}{|S_A \cup S_B|} \tag{1}$$

この方法を用いて全てのウェブページ間の類似度を計算することは不可能ではないが、大規模なデータセットに対して全てのパターンの類似度を計算することは現実的ではない。そこで我々は、Locality-Sensitive Hashing(LSH)[?] を適用することで計算量の削減を行った。

3.3　実験

提案手法の実装は主に Python 言語を用いて行った。機械学習の SVM には LIBSVM[3]、ランダムフォレストには Orange[4]を利用した。計算の実行環境は、OS が Linux、CPU が Intel(R) Core(TM)2 Quad CPU Q6600 の 2.4GHz、メモリが 6GB である。データベースには Postgresql の 8.3 を利用した。

3.4　データセット

実験に用いるデータセットは広く公開されており、関連研究との比較が可能なものが望ましい。そのような条件を満たすデータセットとして、WEBSPAM-UK2007 を用いた。WEBSPAM-UK2007 はウェブスパム検出の研究のために Yahoo! Research Barcelona において公開されているもので、Universitá degli Studi di Milano において.uk ドメインを対象にクロールされたものである。ウェブページは WARC（Web ARChive）フォーマット[5]を用いてアーカイブされており、WARC フォーマットを扱うために Universitá degli Studi di Milano において公開されている LAW Library[6]を用いた。

データセットは 114,529 のウェブサイトから成る 105,896,555 のウェブページを含んでおり、圧縮された状態で 560GB の膨大なデータ量である。我々は各ウェブサイトのウェブページ数の上限を 400 とした要約バージョン（46GB）を実験に使用した。

データセットの一部のウェブサイトには有志によりラベルが付与されている。ラベルは 0〜1 の値をとり、0.5 より小さな値はウェブスパムではないウェブサイト（ノンスパム）を示し、0.5 より大きな値はウェブスパムを示す。0.5 はボーダーラインを示す。表 1 にラベルの分布を示す。このラベルのうち、2/3 が訓練用データ、1/3 がテスト用データとして機械学習用に提供されている。本研究ではこの訓練用データとテスト用データを使って実験と評価を行った。ラベルが undecide 若しくは borderline となっているものは評価の対象から除外した。

Table 1　WEBSPAM-UK2007 データセットにおけるラベルの分布

ラベル	ラベル数	割合
Spam	344	5.3%
Non-spam	5709	88.1%
undecided or borderline	426	6.6%
合計	6479	100%

3.5　機械学習による評価

機械学習による提案素性の評価実験について述べる。機械学習の手法には SVM とランダムフォレストを用いた。

評価尺度には AUC(Area Under Curve) と F 値を用いた。AUC は横軸の尺度を False Positive Rate、縦軸の尺度を True Positive Rate とした ROC カーブの曲線下面積を求めたものである。AUC と F 値の計算は PERF[7]を用いて行った。

$$TruePositiveRate = \frac{TP}{TP + FN} \tag{2}$$

$$FalsePositiveRate = \frac{FP}{FP + TN} \tag{3}$$

[3]LIBSVM, http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[4]Orange, http://www.ailab.si/orange/
[5]http://bibnum.bnf.fr/WARC/warc_ISO_DIS_28500.pdf
[6]LAW Library, http://law.dsi.unimi.it/software/
[7]http://kodiak.cs.cornell.edu/kddcup/software.html

データセットのウェブサイトは 275 個の素性が予め計算されている。275 個のうち 96 個は内容ベースの素性で 179 個はリンクベースの素性である。ここで内容ベースの素性を C、リンクベースの素性を L、提案素性を Sims とし、C+L をベースラインの検出器、C+L+Sims を提案手法として評価を行った。

機械学習はサポートベクターマシン (SVM) とランダムフォレスト (RF) を用いた。データセットのスパムとノンスパムの割合が約 1:17 とアンバランスであるため、訓練に用いるノンスパムの数を調整することでスパムとノンスパムの割合を変化させて学習を行った。この結果より、提案素性を加えることで検出器の精度が向上することを確認した。また、AUC が最も高かった RF の 1:5 の ROC カーブを図 1 に示す。この ROC カーブから、提案素性では False Positive Rate が 0 付近の状態で True Positive Rate が 0.4 であることが分かる。これはノンスパムをウェブスパムと誤検出することなく、4 割近いウェブスパムを検出できることを示しており、誤検出が重大な問題となるウェブスパムの検出において重要な改善である。

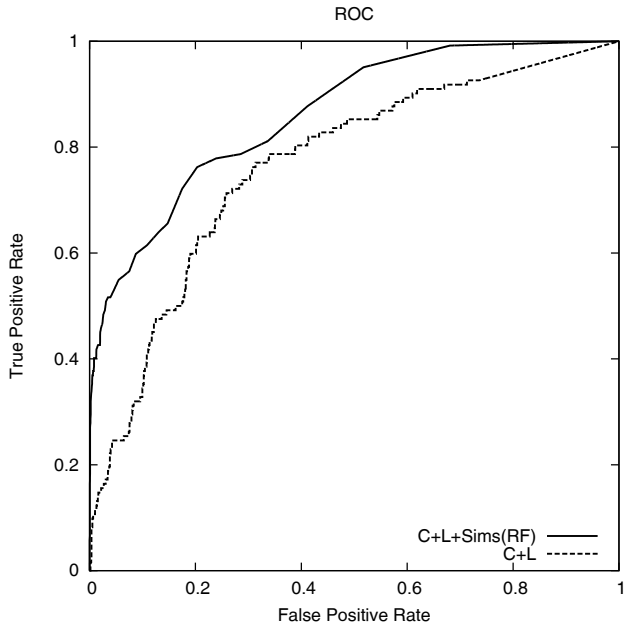

Fig. 1　ROC カーブ

3.6　関連研究との比較

関連研究との比較として AIRWEB2008 の結果との比較を行った。AIRWEB2008 はウェブスパム検出のワークショップで、実験に使用した WEBSPAM-UK2007 データセットが用いられている。表 2 に AIRWEB の結果との比較を示す。この結果より、我々の検出器は AUC と F 値の両尺度において比較的良好な結果を示しており、提案素性の有効性が確認できる。

Table 2　AIRWEB2008 との比較

	AUC	F 値
Geng et al.	0.848	0.470
Tang et al.	0.824	0.368
Abernethy and Chapelle	0.809	0.359
Siklosi and Benczur	0.796	0.318
Bauman et al.	0.783	0.257
Skvortsov	0.731	0.243
C+L+Sims(SVM)	0.823	**0.564**
C+L+Sims(RF)	**0.859**	0.533

4　まとめ

自動化された手法により生成されるタイプのウェブスパムを検出することを目的として検討を行った。そのために、ウェブページ間の構造に着目した *tags* や *scripts* といった素性を構築し、提案素性がウェブスパムの検出に有効であることを WEBSPAM-UK2007 データセットを用いて実証した。また、提案素性を用いて構築した検出器が AIRWEB2008 の結果に対して優れていることを確認した。

発表論文

[1] 北村順平, 青野雅樹 “ウェブサイト間の類似度を用いたウェブスパムの検出”, Journal of the DBSJ, Vol.8, N0.1, pp.143-148, 2009