2008－NL－188 (7)
2008／11／26

ウェブサイト間の類似度を用いたウェブスパムの検出

北村 順平 青野雅樹
豊橋技術科学大学 情報工学系
〒441-8580 豊橋市天伯町雲雀ヶ丘 1-1
{kitamura, aono}@kde.ics.tut.ac.jp

あらまし
ウェブスパムでは、より多くのトラフィックを集めることを目的に様々なスパミング手法が用いられている。ウェブの専門家がウェブスパムを識別することは不可能ではないが、膨大な数のウェブスパムを1つ1つ識別することは非現実的である。そこで我々は、機械学習を用いることで半自動的にウェブスパムを検出する手法を提案した。本手法では訓練用のウェブサイトとテスト用のウェブサイト間の類似度を求めることでウェブスパムの検出を行う。類似度は KNN と SVM を組み合わせた SVM-KNN を拡張したものを用いた。WEBSPAM-UK2007 データセット [1] を用いて本手法の評価を行った結果、効率的にウェブスパムを検出できることを確認した。

Web Spam Detection using Similarity of Websites

Junpei Kitamura and Aono Masaki
Dept. of Information and Computer Sciences, Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi 441-8580 JAPAN
{kitamura, aono}@kde.ics.tut.ac.jp

Abstract
Web spams use many kind of techniques to achieve more traffic from search engines. A web specialist can identify a webspam from websites but it's a distant idea to identify all webspams from a huge number of websites. We propose a method which can semi-automatically detect webspams by applying machine learning techniques. Our method uses similarity of websites to detect webspams. Similarities are determined by KNN, SVM and SVM-KNN. Experimental results on WEBSPAM-UK2007 datasets [1] show that we can efficiently identify webspams.

1 はじめに

ウェブは様々な情報を含む巨大なネットワークであり、膨大な数のウェブサイトが存在する。

利益を目的としたウェブサイトの場合、ウェブサイトへのアクセスが利益に直結するため、より多くのアクセスを集めることがより大きな利益へと繋がる。ウェブサイトへのアクセスの多くは検索エンジン経由だが、ユーザは検索結果の上位2,3個のウェブサイト以外は滅多に訪問しないため、検索結果の表示順位をあげるために SEO (Search Engine Optimization) 等の手法が頻繁に用いられる。しかし、一部のウェブサイトは、ウェブページに大量のキーワードを埋め込む等の SEO の範囲を超えた行為により、本来のウェブサイトの価値以上の表示順位を獲得している。このような行為をスパミングと呼び、ウェブサイトのことをウェブスパム、行為者のことをスパマーと呼ぶ。

E-mail におけるスパムの場合、例えば、単

語別にスパム率の概念の与えるベイジアンフィルタにより高い精度でスパムの検出が可能である。その理由は、E-mail におけるスパムはメール本文に現れる単語が特徴的であることが多いからである。

しかし、ウェブサイトの場合、特定の単語が出現したことを理由にウェブスパムと判別することはできないため、E-mail におけるスパム検出とは異なるアプローチが必要となる。例えば、“バイアグラ”という単語を含む E-mail はスパムの可能性が高いが、同じことをウェブサイトには適用できない。バイアグラを扱う通販サイトの場合、ウェブスパムとは言い切れないからである。

本研究は、ウェブサイトの内容に基づく特徴量からウェブサイト間の類似度を求めることでウェブスパムの検出を行う。

2 ウェブスパムの種類

ウェブスパムの検出に関する研究は、主に内容ベースのスパム、リンクベースのスパムに大別できる [2]。本研究は、内容ベースの特徴量を用いてウェブスパムの検出を行うため、内容ベースのスパムについて説明する。

2.1 内容ベースのスパム

検索エンジンは、検索クエリに適合するウェブページを返すシステムである。その際、ウェブページ内の単語の重み付けを行うために TF-IDF(Term Frequency - Inverse Document Frequency)が用いられる。内容ベースのスパムは、2.2 に説明する手法を用いて TF-IDF のスコアに働きかけるものである。TF-IDF の TF は単語 t のウェブページ内の出現頻度、IDF は単語 t を含むウェブページの数に対応するスコアである。あるウェブページ p における単語 t の TF-IDF スコアは下記の式を用いて求められる。

$$TFIDF(p)=\sum_{t} TF(t)\cdot IDF(t)$$

IDF は単語 t を含むウェブページが多いほど小さな値となるため、一般的な単語は重要度が下がり、特定のウェブページにしか出現しない単語の重要度は高くなる。

2.2 内容ベースのスパム手法

スパマーが検索エンジンからより多くのトラフィックを獲得するための手法は、以下の 2 つに大別される。1 つはウェブページの内容と関係のない単語を大量に埋め込み、多くの単語について 0 より高い TF-IDF スコアを得る手法である。もう 1 つは、単語を意図的に繰り返すことで特定の単語に対して高い TF-IDF スコアを得る手法である。

上記の手法は、meta タグやタイトル、アンカーテキストなど、ウェブページ内の様々な箇所で用いられる。

3 提案手法

本手法では内容ベースの特徴量を定義し、それらを用いて機械学習を行うことでウェブスパムの検出を行う。特徴量については Ntoulas [3]、Castillo [4] が定義する特徴量の一部を本手法において利用する。

定義する特徴量を 3.1、用いる学習法を 3.2 において説明する。

3.1 特徴量

以下の特徴量をウェブサイトのホームページとウェブサイトの複数のウェブページに対して求める。

3.1.1 ウェブページの単語の数と平均長

ここでは、ウェブページの可視テキストの単語数と単語の平均的な長さを求める。

3.1.2 ページタイトル内の単語数

ここでは、title タグ内の単語数を求める。検索エンジンは title タグ内の語を重要視するため、title タグに大量のキーワードを使用するスパムの場合、単語数が多くなる。

3.1.3 アンカーテキストの割合

ここでは、ウェブページの可視テキストに対するアンカーテキストの割合を求める。title タグと同様にアンカーテキストに大量のキーワードを用いるスパムが考えられる。

3.1.4 可視テキストの割合

ここでは、ウェブページのソースコードに対する可視テキストの割合を求める。

3.1.5 圧縮率

ここでは、可視テキストの圧縮率を求める。キーワードの繰り返しや、内容の繰り返しによるスパムの場合、圧縮率が高くなる。

3.1.6 クエリの再現率と適合率

ここでは頻繁に用いられる検索クエリに対するウェブページの再現率と適合率を求める。

3.2 機械学習

KNN (K 近傍法) と SVM (サポートベクターマシン) を組み合わせた SVM-KNN を用いて機械学習を行う。SVM-KNN は KNN と SVM を単独で用いた場合に比べ、良い結果が期待される [5]。SVM-KNN は、KNN の近隣オブジェクトが同一クラスのときは KNN を用いて分類を行い、それ以外のときは SVM を用いて分類を行う。これらの機械学習を行うために、3.1 の特徴量を多次元のベクトル空間にマッピングする。

4 実験

4.1 データセット

Yahoo! Research においてウェブスパムの研究用途に公開されている WEBSPAM-UK2007[1]を用いて実験を行った。このデータセットは、2007 年 5 月に.uk ドメインを対象にクローリングされた 114,529 のウェブサイト、105,896,555 のウェブページから成る。6479 のウェブサイトが人手によってラベル付けされており、その 2/3 が訓練データ、1/3 がテストデータとして提供されている。本研究はこれらのラベル付きデータを用いて機械学習を行った。

4.2 スパムの定義

WEBSPAM-UK2007 データセットにおいて、以下のウェブページを含むウェブサイトはスパムとして定義される。

- 大量のキーワードを含むページ。
- 自動生成された文章や多くのミススペリングを含むページ。
- 広告のみのページ。
- 関連のないウェブページへの自動的なリダイレクトを含むページ。
- 関連のないウェブサイトへのリンクを含むページ。
- ウェブスパムからのハイパーリンクを獲得しているページ。

データセットは、上記のガイドラインに従ってラベル付けがなされる。データセットは 1 つのウェブサイトに対して複数人によってラベル付けがなされており、0 をスパムでないウェブサイト、1 をウェブスパムとしてラベルを定義している。しかし、SEO とスパミングとの区別が難しいウェブサイトなどは、人によって判断が異なってしまうため、0 と 1 の間の値がラベルとして用いられる。

4.3 評価尺度

表 1: 対応表

		ラベル	
		Spam	Non-spam
予測	Spam	TP	FP
	Non-spam	FN	TN

機械学習の評価に AUC(Area Under Curve)と F-measure を用いる。ラベルと予測値の関係を表 1 に記す。TP はラベルが Spam、予測も Spam で正しく分類されるケースである。同様に TN はラベルが Non-spam、予測も Non-spam で正しく分類されるケースである。FP はラベルが Non-Spam、予測が Spam のケース、FN はラベルが Spam、予測が Non-Spam のケースである。これらの値を用いて F-measure を求める。

$$Recall = \frac{TP}{TP + FN}$$

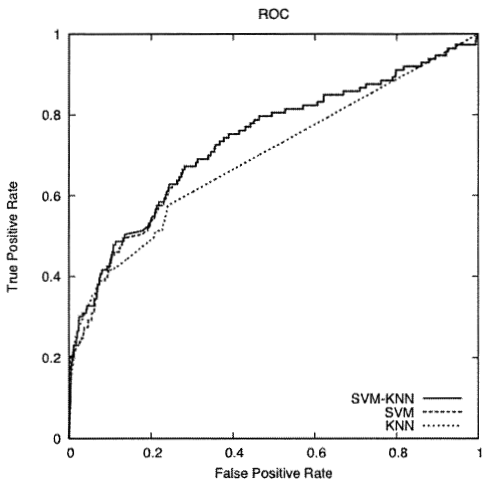

図 1: SVM-KNN, SVM, KNN の ROC カーブ

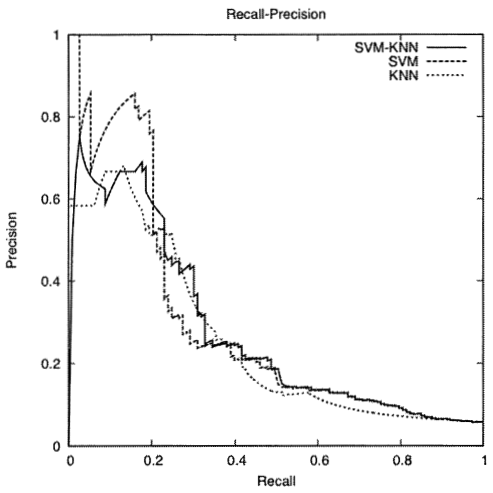

図 2: SVM-KNN, SVM, KNN の Recall-Precision

$$Precision = \frac{TP}{TP + FP}$$

$$F\ measure = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

4.4 実験結果

表 2: SVM-KNN, SVM, KNN に対する評価値

	SVM-KNN	SVM	KNN
AUC	0.736	0.733	0.696
F-measure	0.342	0.293	0.313

KNN を用いる場合、考慮する近隣オブジェクトの数 K を求める必要がある。本実験では、交差検定により適当な K の値として 3 を選んだ。また、オブジェクト間の距離をユークリッド距離により定義した。SVM は LIBSVM [6]を用いて学習とテストを行った。カーネルには RBF を選んだ。

KNN の学習に用いたラベルは 0ー1 の値をとるため、SVM-KNN において SVM と KNN のどちらを適用するか明確ではない。そのため、KNN により得られた予測値 p が p<0.1、若しくは p>0.9 のときは KNN を用いて分類を行い、それ以外の曖昧な予測値が得られたときは SVM を用いて分類を行った。

KNN、SVM、SVM-KNN から得られた実験結果を AUC、F-measure により評価したものを表 2 に示す。また、実験結果を ROC カーブ、Recall-Precision にプロットしたものを図 1 と図 2 に示す。この結果より、SVM-KNN を用いることで、それぞれの手法を単独で用いたときに比べ、分類の精度が向上したことが伺える。

4.5 他の研究との比較

表 3: SVM-KNN, Geng, Skvortsov に対する評価値

	SVM-KNN	Geng	Skvortsov
AUC	0.736	0.848	0.731
F-measure	0.342	0.440	0.140

本実験のデータセットとして用いた WEBSPAM-UK2007 は AIRWEB 2008 の Web Spam Challenge のデータセットとして用いられたものである。我々はこの Web Spam Challenge に参加していないが、比較のために最も F-measure の高かった手法と最も低かった手法の実験結果を引用する。

最も結果の良かった手法は Geng らによるものである [7]。この手法では、内容ベースの特徴量に加え、ウェブグラフに関する特徴量

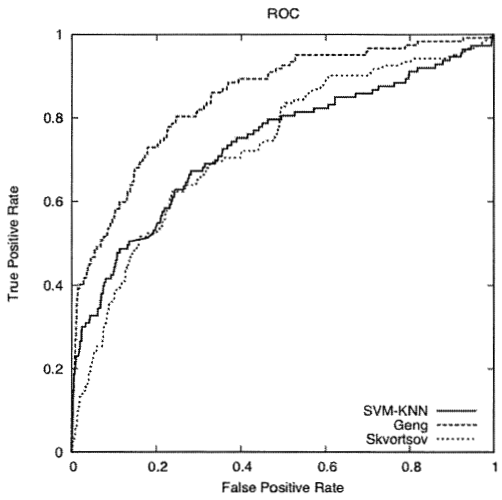

図 3: SVM-KNN, Geng, Skvortsov の ROC カーブ

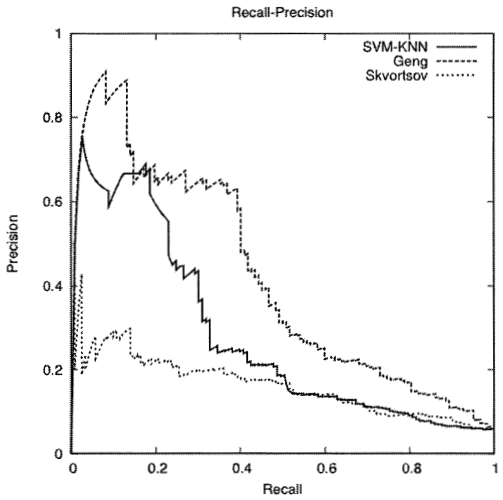

図 4: SVM-KNN, Geng, Skvortsov の Recall-Precision

とホストグラフに関する特徴量を用いて機械学習を行ったものである。もう 1 つは Skvortsov らによるものである [8]。

SVM-KNN とこれら 2 手法の実験結果に対する AUC、F-measure を表 3 に示す。また、実験結果を ROC カーブ、Recall-Precision にプロットしたものを図 3 と図 4 に示す。この結果より、本手法は Geng らの手法に比べ AUC、F-measure の値が低いことが伺える。これは我々の手法と Geng らの手法のおいて用いている特徴量が異なるためである。我々の手法では、内容ベースの特徴量のみを用いたため、リンクベースのスパムを効率良く検出することができなかった。

5 結論と今後の課題

内容ベースの特徴量と SVM-KNN を用いたウェブスパムの検出手法を提案した。実験結果より、ウェブスパムの検出にウェブサイト間の類似度を用いることの有効性を確認した。また、SVM-KNN を用いることで、SVM と KNN を単独で用いたときに比べ分類精度を向上できることを確認した。

本手法では、内容ベースの特徴量のみを用いたため、リンクファーム等のスパムを検出することができなかった。そのため、内容ベースの特徴量に加え、リンク構造を考慮した学習により分類精度を向上させることがこ今後の課題である。

6 参考文献

[1] C. Castillo, K. Chellapilla, and L. Denoyer. Web Spam Challenge 2008. In Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (2008)

[2] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web (2005)

[3] A. Ntoulas, M. Najork, M. Manasse, D. Fetterly. Detecting Spam Web Pages through Content Analysis. In Proceedings of the World Wide Web conference (2006)

[4] C. Castillo, D. Donato, A. Gionis, V. Murdock and F. Silvestri. Know your neighbors: Web Spam Detection using the Web Topology. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 423 - 430 (2007)

[5] H. Zhang, A. Berg, M. Maire and J. Malik. SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2, pp. 2126 - 2136 (2006)

[6] C. Chang and C. Lin. LIBSVM: a library for support vector machines (2001)

[7] Guang-Gang Geng, Xiao-Bo Jin, Chun-Heng Wang. CASIA at Web Spam Challenge 2008 Track Ⅲ. In Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (2008)

[8] Evgeny Skvortsov. Spam Detection via Constraint Programming. In Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (2008)