リンク情報に基づいたウェブスパム検出

佐藤 智博† 青野 雅樹‡

†豊橋技術科学大学情報工学課程 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

‡豊橋技術科学大学情報・知能工学系 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

E-mail: †tomohiro@kde.cs.tut.ac.jp, ‡aono@tut.jp

あらまし 現在，不正にウェブページの価値を高め，検索エンジンをかく乱するウェブスパムの対策が急務となっている．本研究では，ウェブスパムにおけるリンク情報の特異点に着目し，HTML 文書中のアンカータグより得られるリンク情報に基づいた素性を提案する．提案素性は，サイト内アンカータグ数に関する素性，リンク先 URL 長に関する素性, ホスト名とリンク先ホスト名との編集距離に関する素性を含む計 7 素性である. 評価実験は SVM, Random Forest を用い，AUC により評価を行った．特に Random Forest では素性選択を行うことにより検出精度向上と提案素性の重要度算出を行った．評価実験の結果，データセット提供元のワークショップ参加者最高成績を上回ることができ，提案素性の有効性を確認した.

キーワード ウェブスパム，リンク情報，素性選択

Web spam detection based on link information

Tomohiro SATO† Masaki AONO‡

† Department of Information and Computer Sciences, Toyohashi University of Technology
1-1 Hibarigaoka Tenpaku-cho, Toyohashi, Aichi, Japan

‡ Department of Computer Science and Engineering, Toyohashi University of Technology
1-1 Hibarigaoka Tenpaku-cho, Toyohashi, Aichi, Japan

E-mail: †tomohiro@kde.cs.tut.ac.jp, ‡aono@tut.jp

Abstract It is commonly observed that some types of Web spams have peculiar characteristics in their link information. In this research, we focus on anchor tags in HTML documents to detect Web spam. We defined seven additional features including the number of anchor tags, the length of link URL, and Levenshtein distance between its host name and link host name. We adopted SVM and Random Forest as classifiers for Web spam detection. With Random Forest, we attempted to enhance the accuracy of Web spam detection by feature selection based on Gini index. We experimented with WEBSPAM-UK2007 dataset available from Web Spam Challenge 2008. For evaluation measure, we used AUC (Area Under Curve). Our methods outperformed the best team participated in Web Spam Challenge 2008.

Keyword Web spam, Link information, Feature selection

1. 研究背景

インターネットの普及と共に，ウェブページは増加の一途をたどっている．一般にウェブページへのアクセスは検索エンジンを経由したものが多く，検索エンジンで評価され，ランキング付けされた検索結果は，ウェブページへのアクセス数に直結する．利用者にとって, 検索結果上位のウェブページは一般的に有用で, 目的の情報を効率的に発見できる．他方，ウェブページ作成者にとっては，検索結果上位に表示されることでより多くの閲覧者を集め，社会へ広く認知させることができる．このランキングを高めるため，検索エンジンに対し最適化を行う，検索エンジン最適化(SEO: Search Engine Optimization)が行われている．

しかし, 過剰な SEO によりランキングを不正に得るウェブスパムが横行している．その手法は種々あり，ウェブページへの無意味な単語列の埋め込みや URL 長を長くする行為，複数のウェブサイトが相互にリンクを張るリンクファームの構築などである．このようなウェブスパム行為により，検索エンジン利用者にとって有用なウェブページの発見が困難になることや検索結果の信頼性が低下することが懸念される．また，インターネット広告を利用したウェブスパムも散見される．インターネット広告は，クリックやリンク先で

商品が購入されると広告設置者に報酬が支払われる仕組みになっており，このインターネット広告のみで構成されているウェブページはウェブスパムに他ならない．

本研究は，リンクファームやインターネット広告が持つリンク情報の特異点に焦点を当てたウェブスパム検出のための素性を提案するものである．

2. 関連研究

ウェブスパムに関する研究は2000年頃から特に盛んに行われてきた[12][13]．また，ウェブスパム検出を目的としたワークショップである Web Spam Challenge [14]も催され，大規模なデータセットの公開により，研究環境も整ってきた．

ウェブスパムに対するアプローチは様々であるが，ウェブスパム検出には，ウェブページをモデル化する特徴量として「素性」の定義が肝要である．以下に代表的な素性を列挙する．

- **ウェブページ内の単語数**

簡単に実行できるため，ウェブスパム行為にしばしば用いられる方法として，ウェブページ内に大量の単語を埋め込む行為がある．単語数が増えるほど，ウェブスパムである可能性が高い[1]．

- **単語の平均長**

ユーザの検索クエリへの誤入力を狙い，単語を複数繋げるなどしたウェブスパム行為がある．単語の平均長が長いほどウェブスパムである可能性が高い[1]．

- **ウェブページの圧縮率**

ウェブページ内の文章を複製して単語数を増やすことで，検索エンジンから高評価を得ようとするウェブスパム行為がある．こういったウェブページを GZIP で圧縮すると，多様な文章が書かれているウェブページより圧縮率が高くなる傾向にある[1]．

- **コーパス内の頻出単語に対する適合率，再現率**

コーパス内の単語を複数選択し，その適合率と再現率を用いることで，辞書から単語を機械的に引用するウェブスパムを検出する[1]．

- **頻出クエリに対する適合率，再現率**

検索エンジンにより発見されるようにするため，機械生成的に頻出クエリを用いてウェブページを作成するウェブスパム行為がある．こうしたウェブページを検出するため，クエリに対する適合率や再現率を用いる[2]．

- **ウェブページの類似度**

機械生成により複数作成されるウェブスパムは HTML のタグ構造やスクリプトコードに類似性が認められため，この類似度を素性とする[3]．

- **TrustRank**

ウェブスパムではないウェブページからウェブスパムへのリンクは貼られないという概念により，PageRank 的手法を用いてページの重要度を算出する[4]．

3. 提案手法

ウェブスパム検出の素性として以下の 7 素性を提案する．

(1) サイト内アンカータグの総数
(2) サイト内アンカータグのページ平均数
(3) リンク先 URL の平均長
(4) リンク先 URL の平均長(URL クエリを除く)
(5) リンク先 URL の平均長 × サイト内アンカータグのページ平均数
(6) ホスト名とリンク先 URL との編集距離の最小値
(7) $\frac{\text{ホスト名とリンク先 URL との編集距離の最小値}}{\text{ホスト名の長さ}}$

素性の算出はウェブページごとではなく，ウェブサイトごとに集約して素性を算出する．また，同一ホスト内へのリンクは除外する．

1.1. サイト内アンカータグ数に関する素性(1)，(2)

図 1 のようなリンクファームではウェブサイト同士がリンクを張り合うため，多数のリンクが存在する．一方，通常のウェブサイトではサイト内のリンクが増えることはあっても，他ウェブサイトへのリンクはそれ程多くはならないと考えられる．

これに基づいて提案する素性が素性(1)，(2)である．素性(1)ではウェブサイト内のアンカータグの総数を算出することでウェブスパム検出を目指している．また，素性(2)ではページ平均数を算出することで，ウェブサイトの規模に左右されないウェブスパム検出を目指す．

1.2. リンク先 URL 長に関する素性(3)，(4)，(5)

インターネット広告を不適切に設置し，利益をあげる行為が行われており，こういったウェブサイトもウェブスパムの一部である．この種のウェブスパムでは，ウェブページ内に意味のあるコンテンツはほぼ無く，インターネット広告のみで構成されていることが多い．

インターネット広告の特徴を調査すると，URL が極めて長い傾向にあった．そこで提案した素性が素性(3)，(4)，(5)である．素性(3)によりリンク先 URL が平均して長いウェブサイトを検出することで，前述のウェブスパム検出を目指した．一方，素性(4)で URL クエリを除いたのは，インターネット広告の URL が長くなる要因として URL クエリが長いことが多いためである．素性(4)は(3)との比較を主な目的として提案した素性である．また素性(5)では，素性(2)，(3)を掛け合わせることにより，インターネット広告を主なコンテンツ

としたウェブスパム検出を目指した.

1.3. ホスト名とリンク先ホスト名との編集距離に関する素性(6), (7)

リンクファームにおいてリンク元ホスト名とリンク先ホスト名が類似していることがある. 図 2 に例を示す. そこで提案する素性が編集距離を用いた素性(6), (7)である.

編集距離は, ある二つの文字列がどの程度異なっているかを示す値である. 例えば, “money”と“monkey”との編集距離を求めると, 編集距離は 1 となる. 一方, “money”と“apple”との編集距離を求めると, 編集距離は 5 となる.

素性(7)では編集距離の最小値を調査対象ホスト名長で正規化することで, 編集距離が小さいウェブサイトをウェブスパムとして検出することを目的とする. 素性(6)は, 正規化しないことで素性(7)との比較を主な目的とする.

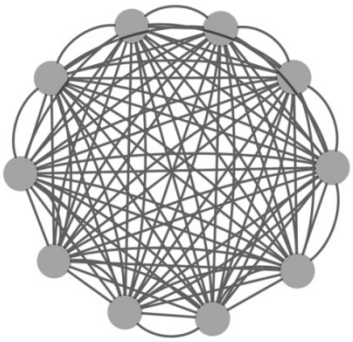

図 1　リンクファーム

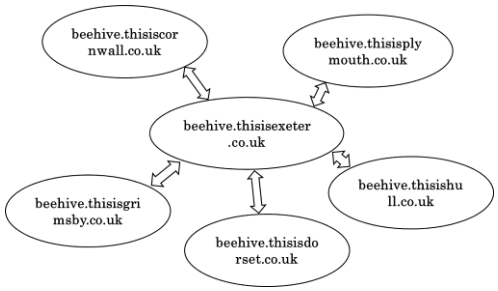

図 2　リンクファームにおけるホスト名の類似例

4. 評価実験

評価実験には次の 2 種類の素性を用意した. 後述するが, データセットにはリンクベース素性とコンテンツベース素性が付属する. 今回提案した素性は, リンク情報に基づいた素性ではあるが, リンク構造を解析した素性では無いため, ベースライン及び提案手法ではコンテンツベース素性のみを用いた.

- **ベースライン(96 素性)**
  コンテンツベース素性
- **提案手法(103 素性)**
  コンテンツベース素性＋提案素性

これらの素性を用いて, 機械学習手法である SVM と Random Forest を用いて実験を行った. 評価尺度は, ROC 曲線とその下面積である AUC である. 実装は R 言語の“kernlab”, “randomForest”及び“ROCR”パッケージを用いて行い, 5 分割交差検定により実験結果を算出している. なお SVM, Random Forest のパラメータはデフォルト値である.

1.1. データセット

データセットは, Web Spam Challenge 2008 ワークショップにより公開されている WEBSPAM-UK2007 である. このデータセットには, 105,896,555 ページから成る 114,529 ホストのデータが収められている. 評価実験では, 各ホストのウェブページ数上限を 400 ページとした要約版を用いた.

またデータセットには, リンクベース素性が 179 素性とコンテンツベース素性が 96 素性付属する. リンクベース素性には PageRank や TrustRank 等が含まれ, コンテンツベース素性にはクエリ適合率, 再現率やウェブページの圧縮率, 平均単語数等が含まれる.

ラベル付けはボランティアにより一部サイトについて行われている. 評価実験ではこのラベルを用いる. ラベルの内訳を表 1 に示す. なお undecided, borderline と判定されたウェブサイトは実験では用いない.

表 1　ラベルの内訳

ラベル	ラベル数
spam	344
nonspam	5709
undecided, borderline	426
合計	6479

1.2. Gini 係数による素性選択

Gini 係数は, 経済の分野では所得格差を表すのによく使われる. 決定木においては, 集団内で復元的にランダム選択された, 任意の二つの要素が異なるクラスに属する確率として捉えることができる[5].

データ集合 D からランダムに選択したデータのクラスが誤分類される確率を Gini 関数と呼び, データがクラス $C_k(k = 1,2,\cdots,K)$ に属する確率$P(C_k)$のとき, 次式で表される.

$$\mathrm{Gini}\big(P(D)\big) = \sum_{i=1}^{K} \sum_{j=1, j\neq i}^{K} P(C_i)P\big(C_j\big)$$

$$= 1 - \sum_{k=1}^{K} P(C_k)^2$$

Gini 係数は素性 A を用いた分割による Gini 関数の差で，

$$\mathrm{Gini}(A) = 1 - \sum_{k=1}^{K} P(C_k)^2 - \sum_{j=1}^{J} \beta_j \left(1 - \sum_{k=1}^{K} P\big(C_{jk}\big)^2\right)$$

と定義される．ここで，β_jは次式で表される．

$$\beta_j = \frac{N_j}{N} (j = 1,2,\cdots,J),\ \ \beta_j \geq 0,\ \ \sum_{j=1}^{J} \beta_j = 1$$

ただし，Jは分割数，Nは分割前のデータ数，N_jは分割 j 内のデータ数，$P\big(C_{jk}\big)$は分割 j 内のデータがクラス C_kに属する確率である．

Random Forest は多数の決定木を用いたアンサンブル学習であるため，生成された決定木から Gini 係数を算出できる．Gini 係数平均減少量が多い素性ほど重要な素性であるといえる．

1.3. 実験結果

1.3.1. ラベル比を変化させた結果

ラベル比に約 1:17 と大きな偏りがみられるため，ラベル比を可変させ実験を行った．表 2 に SVM，表 3 に Random Forest の結果を示す．

表 2　SVM における AUC

spam : nonspam	1:1	1:2	1:3	1:4	1:5
ベースライン	0.804	0.815	0.790	0.792	0.806
提案手法	0.832	0.804	0.812	**0.839**	0.817

表 3　Random Forest における AUC

spam : nonspam	1:1	1:2	1:3	1:4	1:5
ベースライン	0.831	0.816	0.822	0.834	0.808
提案手法	0.852	0.839	0.844	**0.853**	0.841

総じて SVM, Random Forest 共に提案手法の結果が優れており，提案素性の有効性が確認できた．ラベル比変化による結果の相違をみると，SVM，Random Forest 共に 1:4 の時最も結果が良かったが，双方 1:1 の時と大差無かった．また，SVM と Rnadom Forest との比較をすると，どのラベル比においても Random Forest が優れており，また AUC のばらつきも少ない．従って，今回使用した素性においては，SVM より Random Forest の方が優れているといえる．

1.3.2. 素性選択の結果

Random Forest における素性選択の結果を述べる．実験で用いたラベル比は 1:4 とした．

表 4 に提案 7 素性の重要度の順位(103 素性中)を示す．アンカータグ数に関する素性の順位が高く，特に，アンカータグのページ平均数が有効であることが確認できる．次いで効果の高い素性は，リンク先 URL の平均長に関する素性であった．クエリを除いた素性では順位を落としているため，クエリが URL 長を長くする主な要因である，という分析が正しかったものと考えられる．そして最も素性重要度の順位が高かった素性が，「リンク先 URL の平均長×サイト内アンカータグのページ平均数」であった．一方，編集距離を用いた素性の効果は限定的なものであった．

次に素性選択による AUC の評価を述べる．上位 28 素性を用いることで，AUC の最大値 0.864 を得た．上位 28 素性の内，提案素性は 5 素性含まれ，提案素性はおおむね有効といえる．図 3 にこの 28 素性を用いた時の ROC 曲線とベースライン(96 素性)の ROC 曲線との比較を示す．

ベースラインを全領域に渡り上回っており，提案手法が有効であることが確認できる．立ち上がりに注目すると，*False Positive Rate* が 0.05 に満たない時点で *True Positive Rate* が 0.6 を超えており，低い誤検出率で高い検出率を持つことが分かる．

以上から素性選択を行うことで，提案素性の有効性の確認と検出精度の向上が実現できた．

1.3.3. 関連研究との比較

Web Spam Challenge 2008 ワークショップ参加者と同データセットを用いた研究との比較を表 5 に示す．

Web Spam Challenge 2008 参加者の各手法を説明する．Geng らはコンテンツベース素性とリンクベース素性に対し，C4.5 を適用した[6]．Tang らはコンテンツベース素性とリンクベース素性に対し Random Forest を適用した[7]. Abernethy と Chapelle は WITCH とグラフ構造を用いた学習アルゴリズムを SVM により運用した[8]．Siklosi と Benczur は SVM 等複数の検出器を Random Forest を用いて組み合わせた[9]．Bauman らは SVM とナイーブベイズを組み合わせた[10]．Skvortsov はリンクベース素性のみを用いて，制約プログラミングを用いて分類問題を解いている[11].

北村らのウェブサイト間の類似度を用いた研究[3]は，自動化された手法を用いて生成されたウェブスパムを，HTML タグと Javascript のコードを比較することで類似度を算出している．

提案手法の AUC はすべての関連研究に対して上回り，提案手法の有効性が確認できた．

表 4　提案素性の重要度の順位

素性名	素性重要度順位
リンク先 URL の平均長 × サイト内アンカータグのページ平均数	1
サイト内アンカータグのページ平均数	2
サイト内アンカータグの総数	11
リンク先 URL の平均長	16
リンク先 URL の平均長 (URL クエリを除く)	26
$\frac{\text{ホスト名とリンク先 URL との編集距離の最小値}}{\text{ホスト名の長さ}}$	35
ホスト名とリンク先 URL との編集距離の最小値	61

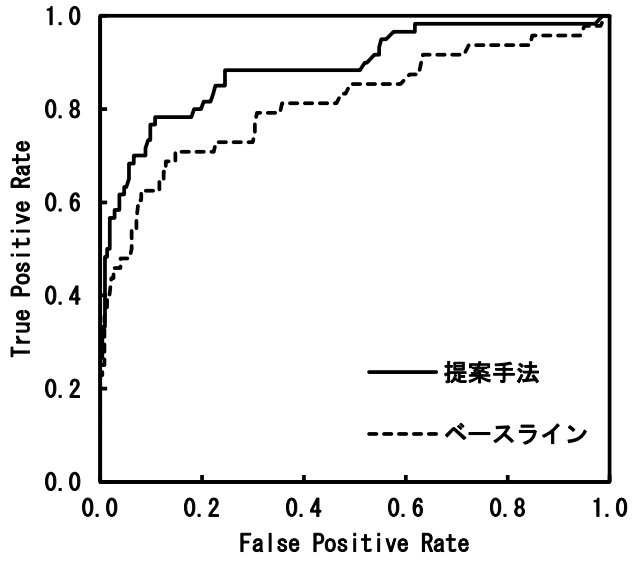

図 3　ROC 曲線比較(Random Forest)

表 5　関連研究との比較

	AUC
提案手法(Random Forest：素性選択)	**0.864**
提案手法(Random Forest)	0.853
提案手法(SVM)	0.839
北村ら	0.859
Geng et al.	0.848
Tang et al.	0.824
Abernethy and Chapelle	0.809
Siklosi and Benczur	0.796
Bauman et al.	0.783
Skvortsov	0.731

5. まとめと今後の課題

本研究では，アンカータグに着目したウェブスパム検出のための素性を提案した．アンカータグより得られる，アンカータグ数や URL 長，リンク先ホスト名との編集距離を用いた 7 素性を用いることで，検出精度向上を得られた．また，素性選択を行うことで，提案素性の有効性と検出精度向上が確認された．

一方で，編集距離を用いた素性は効果が限定的であった．この理由として考えられるのは，ホストを分散している場合である．ホストの中には，“www1.abcdefg.com”と“www2.abcdefg.com”のようにホストを分散している場合があり，これらの編集距離は 1 となるが，同一組織が所有しているサーバである．また，ブログサービスのように同一ホストではあるが，パスによりユーザを分類しているサービスも存在する．これを解決するには，同一組織が所有しているサーバと思われる場合，それらを同一ホストと見なす必要がある．また，ブログサービスにおける対策としては，各々のユーザを独立したサイトと見なすことで解決可能であると考えられる．この場合には，ユーザ名間の編集距離を求める必要がある．

また，URL 長を用いた素性の問題点として，URL 短縮サービスの存在である．URL 短縮サービスでは，サービス事業者が目的の URL へのリダイレクトを行うことで，URL の短縮を可能にしている．こうしたサービスを利用しているウェブスパムは提案素性では検出不可能である．こうした場合には，URL 短縮サービス事業者の URL リストを作り，リストに該当する URL では,リダイレクト先の URL 長を素性として用いる必要がある．

アンカータグ数を用いた素性の問題点は，ポータル

サイトや有益なリンク集のように多数の外部へのリンクを持つ場合も検出してしまう点である．こうした問題は，PageRank や TrustRank 等のリンクベース素性と併用することで緩和できると考えられる．

今回検出器を構成するにあたり機械学習手法である SVM と Random Forest を用いたが，これらのパラメータはデフォルト値を用いた．今後は，SVM のカーネルやパラメータ, Random Forest のパラメータを最適化することで，更なる検出精度の向上が可能であると考えられる．

学習，予測のためのラベル比の差も問題と考えられる．AUC には現れないが，ラベル比が極端であると，一方のクラスに過学習されてしまう．本研究においては nonspam を多く検出する検出器が構築されてしまうという問題が発生する．これを解決するには，世界中のウェブサイトにおけるラベル比を調査し，等しい比で学習することで，現実のウェブスパム状況に即した結果を得ることができると考えられる．

実験用のデータセットとして.uk ドメインを対象にクロールした WEBSPAM-UK2007 を用いた．今後は，他ドメイン，あるいは複数ドメインのデータセットに対して実験を実施することで，提案素性の有効性を検証する必要があると考える．

参 考 文 献

[1] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In Proceedings of the 15th international conference on World Wide Web, pp. 83–92, (2006).

[2] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. You're your neighbors: web spam detection using the web topology. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 423–430, (2007).

[3] 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. 第 188 回自然言語処理研究会, pp. 45‐50, (2008).

[4] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. Link-based characterization and detection of web spam. In Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web, (2006).

[5] 元田 浩, 山口 高平, 津本 周作, 沼尾 正行. データマイニングの基礎, オーム社, pp. 130-138, December (2006).

[6] G. G. Geng, X. B. Jin, and C. H. Wang. Casia at web spam challenge 2008 track iii. In AIRWEB 2008, (2008).

[7] Y. Tang, Y. He, and S. Krasser. Report for webspam 2008 classification modeling. In AIRWEB 2008, (2008).

[8] J. Abernethy, O. Chapelle, and C. Castillo. Web spam identification through content and hyperlinks. In AIRWEB 2008, (2008).

[9] D. Siklosi, A. A. Benczur, I. Biro, Z. Fekete, M. Kurucz, A. Pereszlenyi, S. Racz, A. Szabo, and J. Szabo. Web spam hunting @ budapest. In AIRWEB 2008, (2008).

[10] K. Bauman, A. Brodskiy, S. Kacher, E. Kalimulina, and R. Kovalev. Webspam-uk2007 challenge: Data analysis school in moscow. In AIRWEB 2008, (2008).

[11] E. Skvortsov. Spam detection via constraint programming. In AIRWEB 2008, (2008).

[12] The search engine spam police, http://searchenginewatch.com/2159061.

[13] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web, (2005).

[14] Web Spam Challenge – HomePage, http://webspam.lip6.fr/wiki/pmwiki.php.