ウェブページの統計的解析に基づく素性を用いたウェブスパム検出

パオ ワンナー† 青野 雅樹‡
†豊橋技術科学大学 情報工学系 ‡〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1
E-mail: †vanna@kde.cs.tut.ac.jp, ‡aono@tut.jp

あらまし 最近、大量のウェブ情報の中にウェブスパムが埋め込まれ，検索エンジンの信頼性を低下させていることが問題となっている．本研究では，ウェブスパム検出のため，HTML 文書の統計解析を行い，それに基づく素性を提案する．実験では Webspam-UK2007 データセットを使用し，提案素性を導出した．提案素性の妥当性を評価するため，SVM と Random Forest を用いて実験を行った．

キーワード 統計解析，HTML 文書，素性，ウェブスパム検出

Web Spam Detection using Features based on Statistical Analysis of Web pages

Vanna PAV† and Masaki AONO‡
†Department of Information and Computer Science, Toyohashi University of Technology
‡Hibarigaoka 1-1, Tenpaku-cho, Toyohashi-shi, Aichi, 441-8580, Japan
E-mail: †vanna@kde.cs.tut.ac.jp, ‡aono@tut.jp

Abstract In this paper, we describe the statistical analysis to find out what kind of properties are commonly observed to distinguish spam and non-spam web pages by using 4900 web pages from web spam corpora Webspam-UK2007. Based on this analysis, we propose “features” for web spam detection. To evaluate the validity of the proposed features, we have experimented with two classifiers SVM and Random Forest.

Keyword Statistical Analysis, HTML document, Feature, Web Spam Detection

1. はじめに

大量のウェブサイトから希望する情報を取得するため，多くのユーザーは検索エンジンを利用している．ユーザーは，検索結果ページの上位に最も関連性の高いページが現れ，目的の情報を効率的に発見できることを期待する．検索結果ページに表示されないウェブサイトは，多くのユーザーにとっては存在しないのと同じことである．そこで，このランキングを高めるため，検索エンジンに対して最適化を行う，SEO(Search Engine Optimization)が行われている．SEO とは、検索エンジンを対象に，ウェブページへの無意味な特定キーワードの埋め込み，リンク長を長くする行為における検索結果ページの上位に表示されるよう，対策を行う．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．近年，このようなウェブスパムは非常に増えており検索エンジンの質を保つための対策が重要となる．

本研究では，ウェブスパム検出のため，統計的な解析を行い，それに基づく素性を提案することを目的とする．実験では Webspam-UK2007 の約 4900 ホストラベル付けのデータセットを使用し，そこから提案素性を導出する．具体的には HTML 文書とページ内容を用いて，HTML ヘッダー情報，各タグの数，リンクと IMG タグの属性情報およびページ内の単語数を抽出し，これらから提案素性候補を決定する．一方，抽出データからスパムおよびノンスパムのウェブページの統計データを分析し，最終的に提案素性候補から有効な素性を提案する．提案素性の評価実験を行

うため, 機械学習でよく使用される SVM と Random Forest を用いる． 本論文の構成は以下の通りである．2 節と 3 節では，ウェブスパムとその関連研究について述べる．4 節では，データセットを統計解析し、その結果に基づいた有効な提案素性について述べる．5 節では，提案素性の有効性を検証するために，実験を行い，実験結果と考察を述べる．6 節では，まとめと今後の課題について述べる．

２．ウェブスパムとは

Google によると，ウェブスパムとは隠しテキスト，偽装クローキング, 誘導ページなどの手法により, Google のウェブクローラ（web crawlers）を欺こうとするサイトである[3]． Yahoo では，ウェブスパムとは，検索キーワードと十分な関連性がないにもかかわらず，意図的に検索結果に表示されるように操作をしているページを指すと定義している． これと判定された場合，特定または全部のリンク評価を無効にする，特定のページまたはサイト（ドメイン）全体をインデックスから除外する，といった対応がされる．

ここは，ウェブスパムの特徴について述べている．

HTTP ヘッダー：HTTP ヘッダーから取得できる IP アドレスやユーザーエージェント，リファラーなどの情報をもとにキーワードに関係のある情報を満載したページを，特別なホストアドレスで隠して設置する． これが検索エンジンによって処理されると，特定のキーワードに対して，サイトの掲載順位が上昇する[10].

キーワードの埋め込み：ページ内に大量のキーワードを埋め込む手法である．たとえば，背景と同じ色のテキストをページ内に多量に埋め込むものやテーブル内のテキストの色が背景色と同じ場合などがある．また，タイトルやメタタグの中にキーワードを大量に羅列することで，検索エンジンでの順位上昇を狙うものである．

コンテンツベーススパム：コンテンツと無関係に，機械的に生成されるページで，検索エンジンで上位表示を実現するためだけに行なう行為のことである．コンテンツベーススパムは，検索エンジンにおけるランキング付けで用いられる TFIDF に対し最適化を行うことが多い[5].

リンクベーススパム：リンクの数を数えている検索アルゴリズムを欺くため，偽のサイトを膨大に作成し，そのすべてを同じページにリンクする方法である． Google の PageRank や WiseNut など外部からのリンクを重視するエンジンの登場により出現する[2].

ページ転送：これはメタタグ，JavaScript でページを転送することである．人気キーワードでページを上位に表示させ，そこから目的のページを誘導する[11].

3. 関連研究

多く研究では，HTTP 情報やウェブページの内容，ウェブグラフのリンク構造を解析することによるウェブスパムを検出する．

HTTP 情報ベース：HTTP セッション情報，すなわち，IP アドレスと HTTP のセッションヘッダーをホストしているだけに依存するウェブスパム分類を予測する[10].

コンテンツベース：簡単に行えるため，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．ページに含まれ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と想定して検出を行う[5].

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が存在する．こうしたページでは 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為，この類似度を素性とすることでウェブスパム検出を行う[1].

リンクベース：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 などの手法を用いてページの重要度を算出する[2][10].

4. 提案手法

本研究では，ウェブページの統計解析を行い，ウェブスパムを検出できる有効な素性を提案することを目的とする．提案素性は以下の流れのどおりである．まず，HTML 文書から HTML ヘッダー情報, 各タグの数, リンクと IMG タグの属性情報およびページ内の単語数を抽出し，これらから提案素性候補を決定する．そして，抽出されたデータからスパムおよびノンスパムのウェブページの統計データを分析し，最終的に提案素性候補から統計的解析に基づ

く有効な素性を提案する（図 1）.

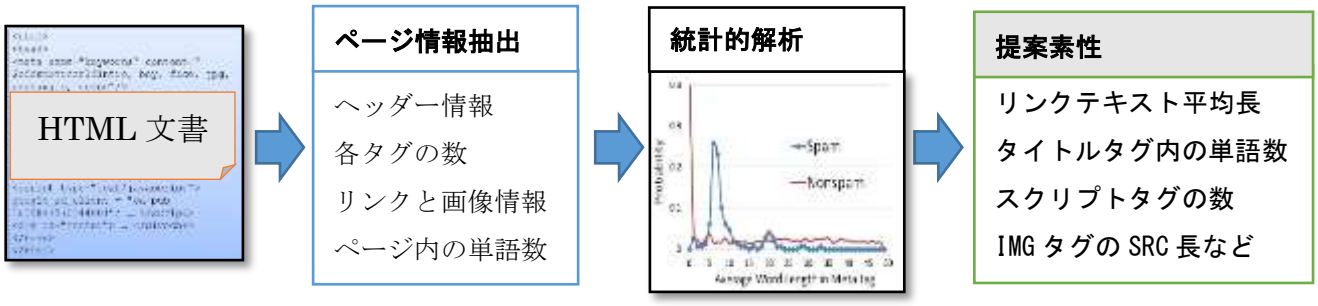


図 1　提案手法の流れ

4.1 ヘッダー情報

本節では，HTTP ヘッダー情報，HTML 文書のタイトル情報およびメタ情報の統計的解析について説明する．

4.1.1 HTTP ヘッダー情報

スパムページがウェブ上でどのようなページが他ページよりもスパムである可能性が高いかどうかを調べる．これを評価するために，ドメインの種類とホストの IP アドレスなどについて実験を行った．図 2 は，スパムページとノンスパムページのホスト IP アドレスの比較を示す．スパムページの IP アドレスは，いくつかの IP アドレス範囲に集中している．具体的には，65.*－90.*と 190.*－215.*に集中し，合計約 80%に占める．スパムとノンスパムページの IP アドレスの分布が異なっているから，ウェブスパムの分類素性として，IP アドレスを使用した．

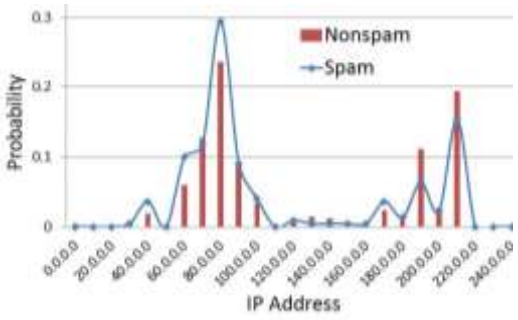


図 2　ホストの IP アドレスの変化

4.1.2 ページのタイトル情報

検索クエリに応じる検索結果の選択時には，検索エンジンは一般的にページのタイトルにクエリキーワードの出現を考慮することがある．いくつかの検索エンジンがページのタイトルに検索クエリの存在のために余分な重みを与える．それに対して，一般的に使用されるスパム技術がページのタイトルにキーワードスタッフィング（Keyword Suffering）という技術が使われている．本実験では，HTML ページのタイトルに含まれるキーワードを着目してスパムページとノンスパムページを統計的に比較し，ウェブページのタイトルがウェブスパム検出のために有効であることを明確する．図 3 では，タイトルの単語数は，スパムとノンスパムページの両方が 6 単語付近に集中しているが，10 単語以上ではスパムページの方が数多く見られる．さらに，スパムページは 60 文字以上の長いタイトルを持つページがノンスパムページより多かったことが実験で分かった．従って，ページのタイトルに含まれる多数の単語および長いタイトルでは，スパムの良い指標であることを観察した．

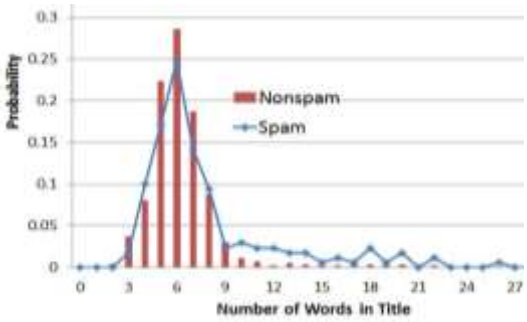
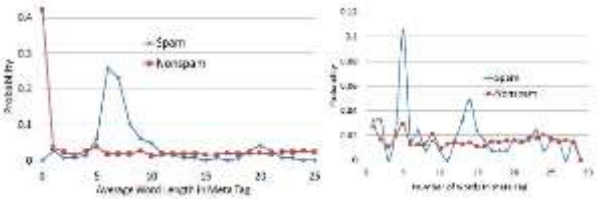


図 3 タイトルタグ内の単語数

4.1.3 メタタグ情報

メタタグは，HTML 内に記述された情報，主に利用者向けの情報ではなく，サーバーや検索エンジン用の情報などが記述されている．Google はメタタグを見てメタタグの Description を Snippet として利用することもあるため，ウェブスパムではよく用いている．本節ではメタタグ内

Keyword 属性及び Description 属性を抽出し，メタタグ内の平均単語長，単語数，メタタグ内のテキスト長について統計的解析を行った．図 4 は，メタタグ情報のスパムとノンスパムページを統計的に比較した結果を示す．スパムページは，(a)メタタグ内の平均単語長が 5 文字から 9 文字付近に集中しているが，ノンスパムページではメタタグ内の keyword 属性をほとんど使用しないことがわかる．また，(b)メタタグ内の単語数では，スパムページが 5 単語及び 15 単語でピーチになり，ノンスパムページに比べると異なっている分布である．

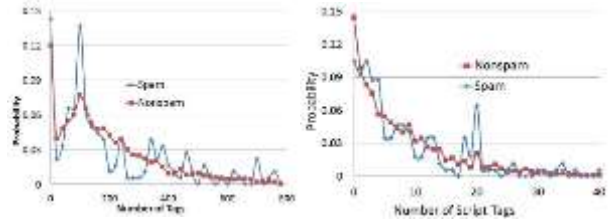


(a) 平均単語長　　(b) 単語数

図 4　メタタグ情報

4.2 各タグ情報

人間は，新聞を見るとき最も重要と判断するのはタイトルであるが、ページの構成にも重要であると考える．次に HTML のタグの重要性に着目していく． 本節では HTML 文書を構築した各タグ数を抽出し，Script タグ，div タグ，td タグ，li タグなどのデータについて統計的分析を行った．図 5 は，HTML 文書のタグ数のスパムとノンスパムページを統計的に比較した結果を表す．各分布は異なっているため，ウェブスパムの分類を使用するには有効であることがわかる．たとえば，SCRIPT タグ数の確率分布についてスパムページがノンスパムページより数多く，20 個でピーチになる．



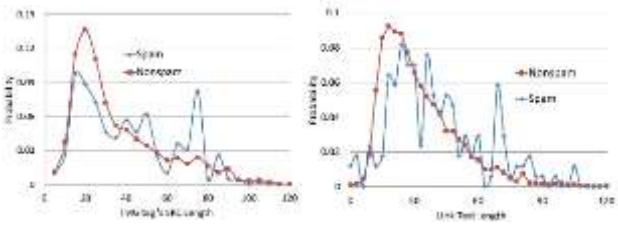
(a)ページ全タグ数　　(b) SCRIPT タグ数

図 5　 HTML 文書の各タグ数

4.3 リンクと画像情報

一般的な検索エンジンは，リンクのターゲットページのコンテンツを記述するため，ページ内のリンクのテキストを考慮することがある．主な考え方は，もしページ A はページ B を指すリンクのテキスト“コンピュータ”がある場合，我々は，ページ B 内にコンピュータというキーワードが無い場合でも，このページ B がコンピュータの話題だと考えている．検索エンジンの場合でも、これを考慮して単語 “コンピュータ”を含む検索クエリに対する結果として，ページ B を返す可能性がある．その結果，いくつかのウェブスパム技術が他のページのリンクのテキストを提供する目的のために存在している．

リンクベースのスパムを調べるために，リンクと画像情報であるリンクの数，リンクの平均長さ，リンクテキストの平均長さ，IMG タグの SRC 長などを計算し，スパムページとノンスパムページを統計的に比較し，ウェブスパム検出のために有効な素性であることを明確する．結果は図 6 に示し，両者の分布が明らかに異なっているとわかる．たとえば，(b) リンクテキストの平均長さについてスパムページの方が長く，70 文字以上のリンクテキストは約 20%もある．



(a) IMG の SRC 長　　(b)リンクテキスト平均長さ

図 6　リンクと画像情報

4.4　ページ内容

ウェブスパムを作成する際，キーワードスタッフィング（Keyword stuffing）が人気な方法であることが知られている．キーワードスタッフィングの技術は，ウェブページの内容に無関係である人気単語などがたくさん埋め込む．単語を混合することにより，スパムページは複数のクエリを照合し，それによってより多くのユーザーに見られる．また，もう一つコンテンツベースのスパム技術は，いくつの単語を連結して長い複合語を形成するものである．例とし

ては,“musicfree”, “downloadvideo”, “onlineshopping”のような単語を持つウェブページがスパムページである可能性が高いと考える．図 7 は，ページの最大単語長についてスパムページとノンスパムページを統計的に比較すした結果を示す．ページの最大単語長についてはスパムページの方が長く，53 文字の最大単語長では約 10%があり，ノンスパムページより 5 倍も多かった．

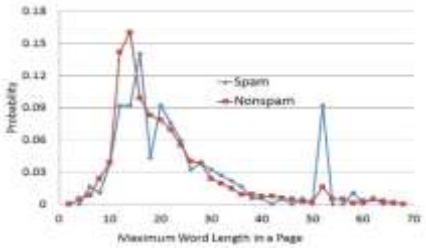


図 7　ページの最大単語長

4.5　提案素性

上記に述べたすべての提案素性をまとめてみると 35 素性があり，その中に独自で考えた素性（下線）が 22 個であった。

表 1 すべての提案素性

種類	各素性
ヘッダー情報	メタの単語数,メタの平均単語長, メタのテキスト長，Domain Type，タイトルの単語数，タイトルの平均単語長，タイトルのテキスト長，ホストの長さ，ホスト IP アドレス
タグ数	tags，　script, div, br, td, li, p
ページ内容	単語数，平均単語長，最大単語長，長さ 1-3 の単語数，長さ 4-9 の単語数，長さ 4 以上の単語数，長さ 10 以上単語数
リンク情報	リンクの数，リンクの平均長さ，リンクテキストの平均長さ，リンクの最大長さ，外部リンクの数，内部リンクの数
画像情報	画像数, IMG タグの SRC 長, IMG タグの SRC の最大長さ

5. 評価実験

本節では，提案素性の妥当性を検証するため評価実験を行った．ウェブスパム検出器の作成は，機械学習手法である SVM と Random Forest を利用した．SVM はパターン認識や回帰分析で多く用いられ，高い汎化性能を誇るため用いた．一方，Random Forest では，Gini 係数による素性の重要度を計算することが可能であるので，提案素性の有効性を検証できるためである．

5.1　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，2007 年 5 月に.uk ドメイン（www.example.*.uk）を対象に，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，ウェブスパムか否か判別が必要であり，WEBSPAM-UK2007 の約 4900 ホストラベル付けのデータセットを使用し，そこから提案素性を導出す．実際にラベル付けデータセットを表 2 に示す．

表 2　WEBSPAM-UK2007 のデータセット

ラベル	ラベル数	割合[%]
Spam	193	3.94
Nonspam	4707	96.06
合計	4900	100

5.2　実験結果

図 8 は，機械学習手法である SVM を用いた実験結果を表す．実験では素性別でスパムページ 1 割ノンスパムページ 5 割（S1:NS5）および全体データ（All-DATA）を行った．この結果をみると，リンクと画像情報ベースの素性は比較的高精度であることがわかる．また，すべての素性を使うと分類精度が約 9 割で高くなる．

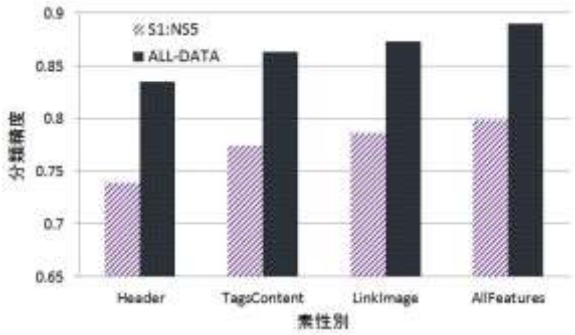


図 8　ウェブスパム分類精度の実験結果

5.3 素性重要度の計算

Gini 係数は，イタリアの数理統計学者 Gini が 1936 年に考案した指数であり， Gini 係数により素性ベクトルの重要度を求めることができる. 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，以下の式で表される.

$$\mathrm{Gini}(\mathrm{P(D)}) = \sum_{i=1}^{K}\sum_{j=1,j\neq i}^{K} P(C_i)P(C_j) = 1 - \sum_{k=1}^{K} P(C_k)^2$$

Gini 係数は素性 A を用いた分割による Gini 関数の差で，

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

と定義される．ここで，β_jは次式で表される.

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

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, $P(C_{jk})$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 による Gini 係数を求めた結果から，リンクの平均テキスト長の素性が最も重要である．リンク情報ベースの素性を用いれば比較的高精度でウェブスパムを分類できると考えられる.

6. まとめと今後の課題

本研究でウェブデータの統計的解析に基づく 35 素性を提案した. 実験では Webspam-UK2007 のデータセットを使用し，そこから提案素性を導出した．リンクベース素性が比較的に高い精度でウェブスパムを検出できることがわかった．また，すべての素性を使用すると，分類精度が約 90%で高くなった.

今後は HTML の色特性，人気キーワード等の素性を追加することや元素性を提案素性と組み合わせることで検出精度を向上させる予定である.

参考文献

[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．日本データベース学会論文誌Vol.8, No.1 (2009).

[2] 佐藤智博，青野雅樹．リンク情報に基づいたウェブスパム検出．DEIM Forum B3-5 (2011).

[3] SEO対策研究所 http://seolab.capoo.jp/ (2012)

[4] WEBSPAM-UK2007. http://barcelona.research.yahoo.net/webspam/datasets/ .

[5] 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).

[6] G. Guanggang, Z. Pengfei, W. Deliang. Web Spam Detection with Feature Fusion. Journal of Computational Information Systems, pp. 1511-1519, (2009).

[7] Y. Liu, R. Cen, Z. Min, S. Ma, L.Ru. Identifying Web Spam with User Behavior Analysis. AIRWeb'08 Proceedings of the 4th international workshop on Adversarial information retrieval on the web, pp. 9-16, (2008).

[8] S. Sahu, B. Donggre, R. Vadhwani. Web Spam Detection Using Different Features. International Journal of Soft Computing and Engineering (IJSCE), Vol.1 Issue-3 (2011).

[9] J. Abernethy, O. Chapelle, C. Castillo. Web spam identification through content and hyperlinks. AIRWeb'08 Proceedings of the 4th international workshop on Adversarial information retrieval on the web, pp. 41-44, (2008).

[10] S. Webb, J. Caverlee, C. Pu. Predicting Web Spam with HTTP Session Information. CIKM'08 Proceedings of the 7th ACM conference on Information and Knowledge management, pp. 339-348, (2008).

[11] T. Urvoy, E. Chauveau, P. Filoche, T. Lavergne. Tracking Web Spam with HTML style similarities. ACM Transaction on the Web (TWEB), Vol.2, Issue-1, (2008).