2B4-02

機械学習によるWebリンクスパムサイト推定

Link Spam Sites Filtering using Machine Learning.

高橋 大和[*1] 熊谷 雄介[*2] 小長井 俊介[*1] 片岡 良治[*1]
Yamato Takahashi Yuusuke Kumagai Syunsuke Konagai Ryouji Kataoka

[*1] NTT サイバーソリューション研究所 NTT Cyber Solutions Laboratories
[*2] 筑波大学 システム情報工学研究科 Graduate School of Systems and Information Engineering University of Tsukuba

The link spam detection is one of the important problems. We proposed a method is based on the link spam detection algorithm using trusted hosts [Gyongyi et. al. 2006] and improved spam link detection accuracy by removing links within the same domain while blog host is regarded as a domain for each author. However, if a site has high spam score close to 1.0, it is difficult to infer that the site is link spammed or not. We proposed a method of filtering spammed sites using machine learning technique on this paper.

1. はじめに

現在は，インターネットの普及が進み，Web 検索サービスの重要性は年々高まっている．Web 検索サービスでは，ユーザが入力した検索キーワードに応じて，検索結果を適切なランキングで提示することが重要である．適切なランキングを行うための指標のひとつとして，PageRank[Page 1998]に代表される，ハイパーリンクによるネットワーク構造を基とした静的な重要度がある．Web 検索サービスの重要性が増すにつれ，検索結果において上位のランキングを獲得するために，大量のページからリンクを張るようなリンクスパミングが日常的に行われている．そこで，このようなリンクスパミングによる影響を排除することは，適切なランキングを行う上で重要な課題である．

リンクスパミングを行っていると推測されるホスト群を Web リンクスパム集合として推定する様々な手法が提案されている．主にリンク構造のみを使う手法として，極大クリークを手がかりに推定する手法[Han 2007]や，スパムホストのブラックリストと非スパムホストのホワイトリストを使ってリンクスパム集合を推定する手法[Wu 2007]，また，政府系ドメインや教育系ドメインに属するホストは，怪しいホストへはリンクを張らないであろうという仮説に基づき，信用度を利用して検出する手法[Gyongyi 2006]がある。

我々は，Gyongyi ら[Gyongyi 2006]が提案した信用度を利用したリンクスパム集合検出を基に、同一ドメイン内リンクの排除とブログサービスからのリンクをブログ著者毎に集約して扱うことにより，ページ重要度のランキングを改善する手法[高橋 2009]を提案している．Gyongi ら[Gyongyi 2006]によれば、スパム指数が 1.0 に近い場合もスパムサイトとしているが，NTCIR における実験では，明確にスパムサイトと判定するのは難しかった．そこで，本稿では，1.0 に近いスパム指数を持つサイトがリンクスパムを受けているかどうかを機械学習を用いて判別する実験を行い，その結果を報告する．

2. リンクスパムサイト検出

2.1 信用度を利用したリンクスパムサイト検出

Gyongi ら[Gyongyi 2006]は，政府系ドメイン(.org)や教育系ドメイン(.edu)に属するホスト(URI 表記におけるドメイン)を信用がおけるとしてホスト単位で信用度を与え，ホスト間リンク情報を基に信用度を PageRank 的手法で配布し集計することで，ホスト毎の信用度ランク(Tr)を算出する．同様に，ページ間リンクをホスト間リンクとして集約し，PageRank 的手法でホストをノードとした重要度であるホストランク(Hr)を求める．これらから，式(1)により，ホスト毎にスパム指数 $\tilde{m}x$ が得られる．

$$\tilde{m}x = 1 - Tr / Hr \quad (1)$$

スパム指数は，ホストランクと信用度ランクが共に高ければ 0.0 に近い値をとり，信用度ランクが 0.0 に近いほど 1.0 に近い値をとる．特に，信用度ランクの値を持つホストからリンクを受けていない場合，つまり，信用度ランクが 0.0 の場合は，常にスパム指数は 0.0 となる．これは，信用が置けるホストが属するネットワークから孤立していることを示す．

しかし，ホスト単位で扱った場合，Web サイトが負荷分散や機能分割のために複数のホストから構成されているときは，複数のホストからランク値を配布されるため，ホストランクが高くなる．また，著者毎にディレクトリを持つブログサービスホストでは，ホストが持つ信用度ランクや他のブログ著者が受けているリンクによって配布される信用度ランク値を別の著者のリンクが配布することになる．

2.2 リンクスパムサイト検出の改良

上記の問題を解決するために，管理主体が同一と考えられる同一ドメイン内リンクを排除すること、また、著者毎にディレクトリを割り当てるサービスホストに関して，ディレクトリを付加した形で著者毎に独立したサイトとみなし，ノードをホスト単位からサイト単位に替え、サイトランクと信用度ランクを計算する改良を行った[高橋 2009]. NTCIR Project[Takaku 2006]で公開されているデータセットである NW1000G-04 を使った実験では，スパム指数が 1.0 かつサイトランクが 10.0 以上であるサイトは，全てリンクスパミングを受けているサイトであるという結果を得た．

2.3 改良手法による実験

我々の実験[高橋 2009]では，スパム指数が 1.0 かつサイトランクが 10.0 以上であるサイトはリンクスパムサイトと推定することができた．そこで，独自にクロールした約 9000 万ページの携帯向 Web ページデータセットに対して，改良手法の適応を行い，スパムサイトの検出を行った．抽出されたサイト数は，クロールさ

連絡先：高橋 大和，NTT サイバーソリューション研究所，神奈川県横須賀市光の丘 1-1，Tel (046)859-8777， Fax (046)855-1730，takahashi.yamato@lab.ntt.co.jp

れていないサイトを含み約 400 万サイトであった. 実験結果として, スパム指数 1.0 かつサイトランクが 100 以上であった 44 サイトは, 全てリンクスパムサイトであった. しかし, スパム指数が 1.0 に近いサイト, 例えば,www.google.co.jp の場合, スパム指数は 0.987 と高めの値であり, スパム指数の値だけでは判別が難しいことがわかった.

3. 機械学習によるWebリンクスパムサイト推定

高いスパム指数を持つサイトがリンクスパムサイトか判別可能になることは, リンク情報を利用したランキング指数を計算する上で重要な課題である. リンクスパムを受けているサイトは, リンク構造になんらかの特徴を持つことは, [Han 2007]からも推測できる. そこで, スパム指数が 1.0 であるリンクスパムサイトと信用がおけると考えられるサイトを教師データとして, 機械学習を用いたリンクスパムサイトの判別実験を行った.

3.1 スパムサイトの学習

独自クロールした約 9000 万ページの携帯向 Web ページデータセットに対して, 改良したスパムサイト検出手法を適応し, 信用がおけるサイトとして, 政府系(.go.jp)サイトである 2911 サイトを正例, スパム指数 1.0 かつサイトランクが 100 を越える 44 サイトを負例として学習を行った.

特徴ベクトルは, サイトランク, 信用度ランク, スパム指数, Becchetti ら[Becchetti 2006]の分析から log(リンク数)とランク値との比率, Castillo ら[Castillo 2007]の分析を基にリンク数の分散などの 18 個の素性を採用した. 特徴ベクトルの一覧を表 1 に示す. 信用度ランクやリンクの分散は, 0.0 になる場合があるため, 0.0001 を加算して計算を行った.

表 1　特徴ベクトル

- サイトランク
- 信用度ランク, log(信用度ランク)
- スパム指数, log(スパム指数), exp(スパム指数)
- log(入リンク), 入リンクの分散
- log(出リンク), 出リンクの分散
- log(入リンク) / サイトランク, log(出リンク) / サイトランク
- 入リンクの分散 / サイトランク, 出リンクの分散 / サイトランク
- log(入リンク) / 信用度ランク, log(出リンク) / 信用度ランク
- 入リンクの分散 / 信用度ランク, 出リンクの分散 / 信用度ランク

学習器は, Linux に R2.9.1 をインストールして利用した. 学習モデルとして, SVM, Random Forest, Logistic regression の三種を使い, その判別精度を測定した. SVM は, Kernel として RBF を用い, σ =0.01~0.1 とした. また, Random Forest の tree size は 100～1000 とした.

3.2 スパムサイト推定実験

スパム指数が 0.9 以上である 855 サイトを対象に, 3 種の学習器を用いて, スパムサイトかどうかの判別を行った. 実験結果として, Random Forest が最も精度が良く, 精度 79.1%, 再現率 82%, F 値 0.801 で最も良かった. 図 1 に, Random Forest における特徴ベクトルの重要度を示す. 本実験では, log(入リンク) / 信用度ランクとサイトランクの重要度が高かった.

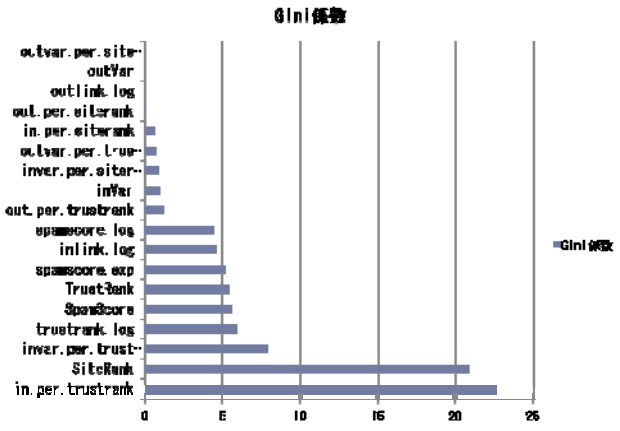

図 1　特徴ベクトルの重要度

4. まとめ

スパム指数が 1.0 であるリンクスパムサイトと信用度の高いサイトを教師データとして学習することで, 0.9 以上の高いスパム指数を持つサイトを 79.1%の精度で判別できることがわかった. 正例, 負例共に特別なラベリングを行わずに用意できるため, Web のような大規模なデータへの適応に有望と考えられる. 今後は, スパムサイト間リンクの除去により, ページ重要度のランキング精度を向上できるか, 実験により検証していく.

参考文献

[Page 1998] L. Page, S. Brin, R. Motwani, and T. Winograd. : The PageRank citation ranking: Bringing order to the web. Technical report, 1998.

[Han 2007] B. Han, M. Toyoda, and M. Kitsuregawa : A Technique for Detecting Web Spam from a Densely Connected Directed Graph of Sites, DEWS2007.

[Wu 2007] B. Wu and K. Chellapilla : Extracting Link Spam using Biased Random Walks From Spam Sets, AIRWeb2007.

[Gyongyi 2006] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen : Link Spam Detection Based on Mass Estimation, VLDB '06.

[高橋 2009] 高橋 大和, 数原 良彦, 小長井 俊介, 片岡 良治 : Web リンクスパム集合検出とスパムリンクを排除したページ重要度, Web インテリジェンスとインタラクション, 電子情報通信学会, 2009.

[Takaku 2006] M. Takaku, K. Oyama, A. Aizawa, H. Ishikawa, K. Minamide, S. Kato, H. Yamana, and J. Hayashi : Building a Terabyte-scale WebData Collection “NW1000G-04” in the NTCIR-5 WEB Task, NII Tech. Rep, 2006.

[Becchetti 2006] 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.

[Castillo 2007] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. : Know 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, 2007.