3P-8

コンテンツベース素性によるウェブスパム検出

佐藤 智博† 青野 雅樹†

†豊橋技術科学大学大学院 工学研究科 情報・知能工学専攻

1 はじめに

ウェブページの急激な増加により，利用者の利益に資する情報を取得する手段として検索エンジンは欠くことのできない存在である．検索結果のランキングは，ウェブサイトへの来訪者の多寡を決定付け，制作者にとって大きな関心事となっている．そこで，ウェブページの検索エンジン最適化(SEO: Search Engine Optimization)を行うことによりランキングの上昇を図っている．

しかし，過剰な SEO により上位ランキングを不正に得る行為が横行しており，これをウェブスパムと呼ぶ．ウェブスパムはリンクベーススパムとコンテンツベーススパムに大別される．リンクベーススパムには，無意味なリンク設置や複数のウェブサイト間で相互にリンクを張り合うリンクファームの形成等があり，コンテンツベーススパムには，脈絡の無い単語の埋め込みや文書の機械生成等がある．

リンクベーススパムに対しては，Truncated PageRank や TrustRank を素性とした検出手法が提案されておりその有効性が示されている．一方，コンテンツベーススパムに対しては，ウェブページ内の単語長や単語数，頻出クエリとの適合率等を素性とした検出手法[1]が提案されている．しかし，文書としての体裁を保っており露骨なスパム手法が適用されていない，製品の宣伝やアフィリエイト目的のウェブスパムでは，単語やトピックに不審な点が少なく前述の手法だけでは検出が難しい．

そこで我々は，ウェブスパムにおける特有の単語遣いを潜在的ディリクレ配分法(LDA: Latent Dirichlet Allocation)[2]の適用によって明らかにし，その単語群との適合率と再現率をコンテンツベース素性とする手法を提案する．例えば，カジノというトピックに対するウェブスパムでは，”カジノ”という単語だけではなく”簡単”や”儲かる”等の利用者が興味を引きそうな単語が一般的なウェブサイトより多く使われていると想定される．こうしたトピック毎の特徴語を LDA の適用により抽出することによって，ウェブスパム検出精度の向上を図る．

Web spam detection using content-based features
† Tomohiro Sato (tomohiro@kde.cs.tut.ac.jp)
† Masaki Aono (aono@tut.jp)
Department of Computer Science and Engineering, Toyohashi University of Technology
1-1 Hibarigaoka Tenpaku-cho, Toyohashi, Aichi 441-8580, Japan

本稿では，提案手法の一連の流れを説明した後，評価実験の結果を示す．2 種のデータセットを用いて実験を行った結果，提案手法の有効性を確認した．

2 提案手法

本節では，提案手法を用いた素性算出法を説明する．その流れを図 1 に示す．

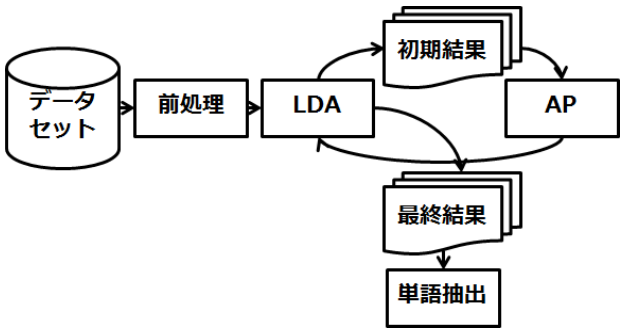

図 1: 提案手法

2.1 前処理

本研究では，英語ウェブページの HTML 文書の本文を抽出対象とする．名詞，形容詞，固有名詞及び未知語に語幹処理を適用し抽出する．その後，LDA の入力のために TF-IDF を Harman 並びに Sparck Jones による正規化によって算出する．

2.2 潜在的ディリクレ配分法(LDA)

LDA は，各文書が潜在的なトピックを持ち，その潜在的なトピックは複数の語から成ると考え，各文書をトピックの確率分布により表現した文書モデルである．

2.1 により得られた単語群に対し LDA を適用する．この時，LDA には予め推定するトピック数を与える必要があるが，ウェブスパム中にどの程度トピックが含まれるか不明であるためトピック数を決定的に指定することが難しい．そこで，LDA に入力するトピック数を推定する方法として Affinity Propagation を用いる．

2.3 Affinity Propagation(AP)

AP は Frey らによって提案されたクラスタリング手法である[3]．AP ではデータ点間で responsibility と availability と呼ばれるメッセージを再帰的に収束するまで計算を行うことで，データ点がどのクラスタの中心(exemplar)に属するかを決定する．AP は

旧来の k-means 等のクラスタリング手法と比較し，初期値依存が無い，類似度の制約が緩い，クラスタ数を自動決定できるといった特徴がある．

本研究では，AP がクラスタ数を自動決定できるという点に着目し LDA に与えるトピック数を推定する．具体的な手順を以下に述べる．(1)十分大きなトピック数で LDA を実行し，初期結果としてトピック群を得る．(2)全トピック間の類似度を Jaccard 係数により算出し AP によってクラスタリングを行う．(3)AP によって得られたクラスタ数をトピック数として LDA を再度適用し最終的なトピック群を得る．

2.5 トピック群からの単語抽出

本研究では，全トピックから n 単語を選択し，それらによって素性を算出する．その指標として TF-IDF における文書をトピックとして算出した値をとる term-score[4]を用いた．

2.5 素性算出

2.4 により得られた単語群によって，検出対象となるウェブページにおける適合率 p 及び再現率 r を算出する．素性単語の適合数 a，文書内の単語数 b と表すと算出式は，

$$p = \frac{a}{n}, \quad r = \frac{a}{b} \tag{1}$$

となる．なお，評価実験ではウェブサイト毎の評価となるため，対象ウェブサイトに属する全ウェブページの平均値を素性とする．

3 評価実験

評価実験では，単語抽出までを Webb らにより提供された Webb Spam Corpus，素性算出・検出実験を Web Spam Challenge 2008 により提供された WEBSPAM-UK2007 を用いた．検出における機械学習手法として SVM 並びに Random Forest を適用した．比較対象には，頻出クエリ及び頻出コーパスに対する適合率と再現率を素性とした研究[1]を用いた．評価尺度には，ROC 曲線とその下面積を採用した．

本実験では，上位 100，200，500，1000 件の頻出クエリ及び頻出コーパスの適合率，再現率の計 16 素性に n=300 とした提案手法の 2 素性を加え評価を行なっている．

図 2 に Random Forest を適用した実験の ROC 曲線を示す．提案手法による素性を加える事で全体的に検出精度の改善がみられる．特に，誤検出が少ない時点での改善が大きく，これは，有用なウェブページをウェブスパムとして誤検出するという問題を回避するという意味で意義が大きいと考えられる．

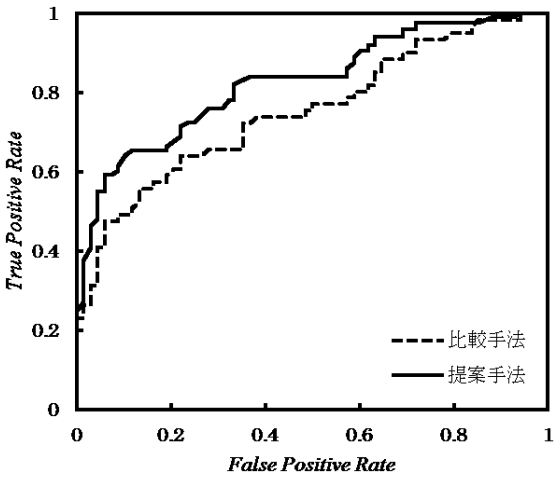

図 2: ROC 曲線 (n=300)

表 1 に素性重要度を算出しその上位 5 件を示す．Random Forest を実行した際に得られる Gini 係数の減少量により得た結果である．

表 1: 素性重要度 (n=300)

順位	素性名
1	**LDA300 単語適合率**
2	上位 200 クエリ適合率
3	上位 500 クエリ適合率
4	**LDA300 単語再現率**
5	上位 200 クエリ再現率

全 18 素性中，提案手法による適合率が最も有用な素性であることを確認した．

4 おわりに

本稿では，LDA によるトピック推定によりウェブスパムにおける特徴語を抽出し，その適合率と再現率を検出時の素性とする手法を提案した．ウェブスパムデータセットによる評価実験の結果，既存手法を上回る検出精度を得た．また，素性選択により，本提案による素性が検出時に高い効果を発揮することが示された．

今後の課題として，異言語での検証，Truncated PageRank や TrustRank 等のリンクベース素性との組み合わせによる検出精度の向上が考えられる．

参考文献

[1] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In Proc of WWW2006, 2006.

[2] David M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent Dirichlet Allocation, The Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[3] B. J. Frey and D. Dueck, Clustering by passing messages between data points, Science, 315, 2007.

[4] David M. Blei, and John D. Lafferty, TOPIC MODELS, Text Mining: Theory and Applications. Taylor and Francis, 2009.