情報処理学会論文誌　データベース　Vol. 3　No. 1　29–37 (Mar. 2010)

# リンク構造とコンテンツを複合的に用いた極少訓練事例によるスプログ検出

吉川 幹人[†1]　佐藤 翔平[†3]
関 和広[†2]　上原 邦昭[†1]

インターネットの普及と情報検索技術の向上によって，ウェブページの商業利用がさかんになり，特定のページへの誘導などを目的としたスパムブログ（スプログ）が大量に作成されている．これらのスプログは，人気のキーワードなどをページ内にちりばめるなどの方法で，検索エンジンの出力結果に不当な影響を及ぼしている．これに対処するため，スプログのコンテンツに着目したテキスト分類手法が数多く提案されてきた．しかし，分類器を学習するためには多数のラベル付けされた訓練事例が必要となり，人的コストが高い．そこで本研究では，ごく少数のブログサイトに対して人手でラベル付けするだけで，多数の訓練事例を得る枠組みを提案する．より具体的には，ブログサイト間のコンテンツとリンク関係に注目することで意味的・リンク構造的に関連するページ集合を同定し，その代表的なページのみにラベル付けすることである程度の大きさの訓練データを得る．先行研究で構築された評価データを用いて実験を行い，本提案手法の有効性を定量的に示す．

# Splog Detection Exploiting Link Structure and Contents Based on Few Labeled Examples

Mikito Yoshikawa,[†1] Shohei Sato,[†3] Kazuhiro Seki[†2] and Kuniaki Uehara[†1]

In the last decade, blogs have grown popular and widely been used as a means to disseminate information by both individuals and organizations. With the growth of blogs, however, the number of spam blogs (splogs) has also been increasing to manipulate the ranking of web search engines, resulted in various problems for users who seek for information on the web. To deal with the problems, there have been several studies for splog detection typically based on supervised classification techniques. While they have been shown effective, a downside of the techniques is that they require manually labeled training data which are costly to create. This paper describes a novel splog detection framework only requiring a few labeled instances. The proposed framework take advantage of both the link structure and the contents of the blogs to identify potential blog/splog clusters and for each cluster nominates a representative page to be manually labeled. Evaluative experiments demonstrate that while significantly reducing the cost of labeling, the proposed framework achieves around 90% of the accuracy obtained with fully labeled data. It is also shown that link structure and blog contents work complementarily for identifying good blog/splog clusters.

## 1. まえがき

現在，インターネット上の情報は急速な勢いで増加しており，百科事典的ウェブサイトWikipedia [⋆1]など，そこには多くの有用な情報が含まれる．その中でも，ウェブというメディアに特徴的なコンテンツとして，ユーザが発信する情報，すなわちUGC（user-generated contents）がある．UGCの例として，個人のホームページ，掲示板，ウェブログ（ブログ）などがある．特にブログは，利用の容易さ，即時性などから，多くのユーザを獲得している．しかし，ブログサイトの増加にともない，商業目的のサイトなどへの誘導やスパイウェアへの感染を目的とした価値の低いスパムブログ（スプログ）と呼ばれるページも大量に生成されるようになった．スプログは，話題のキーワードを無作為に用いたり，人為的に密なリンク構造を構築したりといった方法によって，ウェブ検索エンジンの出力上位に表示されるため，検索精度に悪影響を及ぼす．さらに，スプログ中のコンテンツにアクセスすることでスパイウェアに感染した場合，個人情報の流出や感染したコンピュータの性能低下などの問題も引き起こす可能性がある．

これらの問題に対処するため，スプログの発見・検出に関する研究が行われてきた．たとえばKolariら[1)]はスプログのコンテンツ情報などを用いたテキスト分類の手法により，スプログ検出問題の解決を図った．またLinら[2)]は記事本文に加え，リンク情報，記事タイトルなどに注目したスプログ検出の手法を提案している．これらの手法はいずれも教師付き

†1 神戸大学大学院工学研究科
Graduate School of Engineering, Kobe University
†2 神戸大学自然科学系先端融合研究環
Organization of Advanced Science and Technology, Kobe University
†3 株式会社エヌ・ティ・ティ・ドコモ
NTT docomo, Inc.
⋆1 http://en.wikipedia.org/

の機械学習に基づくため，各ブログサイトがスプログか否かを明示的に示したラベル付き訓練データが必要となる．しかし，これを人手で大量に作成するのは容易ではない．また，そのような訓練データを作成したとしても，スプログのコンテンツ特徴，たとえば話題のキーワードなどは日々変化していくことが予想される．この場合，過去に構築した静的な訓練データが現在のスプログを検出するために有用であるとは限らない．

このような背景から，本研究では，訓練事例作成のコストを大幅に削減するための新しい手法を提案する．提案手法では，ブログのリンク構造とコンテンツの特徴を複合的に利用することで，関連するページ集合を発見する．さらに，発見した各ページ集合を代表するページを同定し，そのページのみに人手でラベル付けを行うことで，同集合に含まれる他のページのラベルに代える．この方法によって，きわめて少数のページを評価するだけである程度の規模の訓練データが作成できるようになり，訓練事例作成のコストが大幅に軽減される．

以下，2 章でスプログの定義とスプログ検出の関連研究について述べる．次に，3 章で提案手法について詳述する．4 章では，提案手法によって発見されたページ集合の評価とスプログ検出の精度について報告する．最後に，5 章で本研究のまとめを述べる．

## 2. 関 連 研 究

### 2.1 スプログとは

スプログは，ブログの形式で作成されるスパムウェブサイトである．Kolari ら[3] は，スプログ作成者の主な動機として次の 2 点をあげている．

(1)　商業目的のサイトへ誘導するためのリンク元とする．

(2)　上記のサイトを含め，提携するサイトが検索エンジンの出力結果の上位に表示されるように不当な操作を行う．

後者は，現在のウェブ検索エンジンが，ウェブページ間のリンク関係を検索結果の順位付けに（一部）用いていることによる．簡単にいえば，多くのページからリンクされているページは検索結果の上位に表示されやすくなる．そこで，大量のスプログを作成し，それらのスプログから提携するサイトへのリンクを作成することで，提携サイトの検索順位を人為的に押し上げることができる[⋆1]．

このような動機から，スプログの記事は，他の正当なブログの複製であったり，意味のない語あるいは文の羅列であることが多い．**図 1**，**図 2** にスプログの例を示す．


**図 1**　リンクの羅列によるスプログの例
Fig. 1　Example of splogs by a link list.


**図 2**　コンテンツの複製によるスプログの例
Fig. 2　Example of splogs by copied contents.

図 1 の例は単純にリンクの羅列であり，見た目からも容易にスプログだと判断できる．一方，図 2 は，やや巧妙なスプログである．普通のブログサイトのように見せかけられているため，一目では判断できないものの，リンクをたどると商業目的のページが多く，スプログであることが分かる．なお，このページの本文は，CNET[⋆2] のレビュー記事の一部を複製したものであった．

以下では，文脈から明らかな場合，記述の簡素化のためにスプログではないブログを単に「ブログ」と呼ぶ．

### 2.2 スプログ検出

Kolari ら[1] は，SVM（support vector machine）を用いてスプログの検出を試み，$F_1$ 値で 0.87 の精度を得た．この研究では，ブログ記事を構成するコンテンツ（語）に加え，ブログ記事が持つ URL やアンカテキストを分類器の素性として提案し，その効果について論じている．また，Lin ら[2] は，投稿時間の分布，内容，リンク先の情報に基づき，ブログ記事の自己相似性を用いてスプログ検出を行う手法を提案し，AUC（area under the ROC curve）で 0.9 以上の精度を得たと報告している．これらの研究の対象が英語圏のスプログであったのに対し，日本語圏における研究としては Ishida[5] の取り組みがある．Ishida は，スプログのリンク構造に着目し，ブログサイトの外部リンクを用いることで，高次元のリンク数においては 9 割以上の精度でスプログの検出が可能であると報告している．

これらの手法によって，ある程度の精度でスプログの検出が可能ではあるものの，従来手

⋆1 PageRank[4] など，実際のアルゴリズムでは，リンク元のページの重要度も考慮する．

⋆2 http://www.cnet.com/

法は静的な訓練データを基にした教師付きの分類手法であり，訓練データに類似したスプログにおいてのみ有効であると考えられる．これに対して，スプログは日々大量に投稿されているうえ，スプログにちりばめられる注目キーワードは時間の経過に従って変化していくと考えられる．このような場合，1 つの解決策として，新しいスプログの特徴を反映した訓練データを再び用意するという方法がある．しかし，訓練データの構築には多大なコストがかかるため，たびたびデータセットを作成し直すというわけにもいかない．よって，大きなコストをかけずに訓練データを作成することができれば，大変有用であると考えられる．

同様の問題に対処する 1 つの枠組みとして，半教師あり学習に関する研究が数多く行われている[6),7)]．半教師あり学習では，少量の訓練データ（ラベル付きデータ）と大量のラベルなしデータを利用することで，少量のラベル付きデータを利用したとき以上の分類精度を達成することができる．代表的な半教師あり学習法としては，EM アルゴリズムを用いた方法[8)] や最小カット法[7)] などがあり，その有効性が知られている．3 章で述べるように，本研究で提案するスプログ検出の枠組みも少量のラベル付きデータを用いるため，その意味で半教師あり学習と見なすこともできる．ただし，本研究は必要なラベル付きデータの数がきわめて少ない点で従来の半教師あり学習とは異なる．また，本研究ではラベル付けすべきページが自動的に検出されるため，能動学習[9)] にも近い．従来の半教師あり学習手法との比較については，4 章で報告する．

### 2.3　本研究のねらい

従来のスプログ検出に関連した研究では，スプログと一般のブログの分類にはコンテンツ（語）が有用であることが知られている．これは，スプログは注目キーワードに加えて，性的なキーワードなど特有の語彙を含むことが多いこと[10)] によると考えられる．しかし，コンテンツのみを利用した方法では，他のブログやウェブページなどの情報源から複製して作成されたスプログを検出することは難しい．一方，スプログの主要な目的の 1 つは，特定のサイトへの誘導である．このため，スプログのリンク先のページは，類似したページ集合である可能性がある．また一般のブログは，内容的に類似したブログどうしでリンクし合うことでコミュニティを形成する傾向があるため，スプログはこのようなコミュニティからは排除されやすいと考えられる．すなわち，リンク構造に注目することで，スプログと一般のブログをある程度区別することが可能であると考えられる．

これらの考察から，コンテンツ特徴とリンク構造に着目することで，スプログとブログのページ群をより正確に分離できる可能性がある．この場合，分離されたページ群に対してラベルを付けるだけで多数の訓練事例が得られるため，低コストでの訓練データ構築が可能になる．すなわち，人手ですべてのページに対してラベル付けする必要がなくなるため，訓練データ作成のコストを大幅に削減できる．

## 3. 提 案 手 法

### 3.1　概　　要

本章では，リンク構造とコンテンツ特徴を用いてスプログとブログをある程度分離することで，分類器生成のための訓練データを低コストで構築する枠組みについて述べる．より具体的には，リンク構造を基にグラフ表現したページ群中に存在する構造が密な部分グラフを複数抽出し，さらにそれらの部分グラフをコンテンツ特徴から二分割する．前者は，リンク関係を持つブログとスプログの集合の抽出を意図しており，後者は，抽出された集合中でのブログとスプログの分離を意図している．よって，最終的に得られたページ集合は，スプログあるいはブログのクラスタを構成するものと期待される．これらのページ集合中の代表的なページのみにラベル付けを行うことで，効率的な訓練データの構築を実現する．

以降，3.2 節でリンク構造を用いた部分グラフの抽出について詳述し，3.3 節でコンテンツ特徴を用いた部分グラフの分割について述べる．さらに，3.4 節で両者を複合的に用いた訓練データ構築の枠組みについてまとめる．

### 3.2　リンク構造によるコミュニティ抽出

前述のとおり，ブログとスプログはリンク構造によって異なるコミュニティを形成していると考えられる．そこで，リンク構造に注目することで，以下のように，ブログおよびスプログの混在したコミュニティ群の抽出を行う．

ウェブページを頂点，リンクを辺とすると，ウェブ空間は巨大な有向グラフととらえることができる[11)]．従来のウェブコミュニティ抽出に関する研究では，このようなグラフから，ウェブコミュニティに特徴的なリンク構造を定義し，その構造を抽出する手法が提案されている．たとえば，Kumar ら[12)] は，ウェブコミュニティを「十分に大きく密な二部グラフ」と定義し，データセットからすべての完全二部グラフを抽出する手法を提案している．一方，Reddy ら[13)] は，DBG（dense bipartite graph）を対応する 2 つのウェブコミュニティと仮定し，コミュニティ群を抽出，さらに同手法を抽出されたコミュニティ間の関連付けに用いた．DBG 内には「ファン」と「センター」と呼ばれる頂点集合があり，ファンはセンターに対して閾値 $p$ 個以上のリンクを持ち，センターはファンから閾値 $q$ 個以上のリンクを持つ．DBG は，シードページからリンクをたどることで抽出することができる．具体的には，まず任意のシードページ $s$ によって集合 $S = \{s\}$，集合 $T = \emptyset$ を定義し，$S$ か

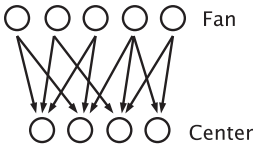



**図 3**　DBG の例
Fig. 3　Example of DBG.

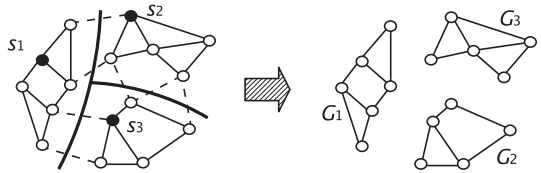



**図 4**　3-way カットの例
Fig. 4　Example of a 3-way cut.

らリンクされているページを $T$ に加え，$T$ をリンクしているページを $S$ に加えることを繰り返し行うことで，密な二部グラフの候補とする．そして，$s_i \in S$ に関して，$T$ への $s_i$ のリンク数が一定値以下なら $S$ から削除する．同様に，$t_j \in T$ に関して，$S$ からのリンクの数が一定値以下なら $T$ から削除することで DBG を得る．**図 3** に，閾値が $p = 2$，$q = 2$ のときの DBG の例を示す．なお，この処理によって通常 DBG は複数抽出される．

本研究で DBG を用いた理由は，コメントスパムなどのスプログの性質から，リンク構造によって抽出した部分グラフ中にはブログとスプログが混在すると考えられるためである．DBG は内部にファンとセンターという二部グラフを仮定しているため，ブログとスプログという異質なデータの混在した部分グラフを表現するために適当だと考えられる．リンク構造に基づくコミュニティ抽出アルゴリズムとしては最大流・最小カット定理による Flake ら[14] の手法など種々提案されているものの，これらのコミュニティはリンク構造の密なページ集合であり，ブログとスプログを含む部分グラフとしてはそぐわないと推測される．

### 3.3　コンテンツ特徴によるグラフ分割

コンテンツ特徴に基づいて類似のページ集合を得るため，まず前処理として各ページを単語ベクトルで表現する．単語ベクトルの各要素には，情報検索で一般的に用いられる TFIDF[15] を用いる．続いて，各ページを頂点 $V$ とする無向グラフ $G = (V, E)$ を構成する．$E$ はページ間の辺であり，重みとしてページ間のコサイン類似度を持つ．なお，どの辺 $e \in E$ も異なる頂点を結ぶものとし，同じ頂点を結ぶ辺（loop）や 2 つの頂点を 2 本以上の辺で結ぶ多重辺はないものとする．

このように構成したコンテンツ特徴によるグラフを基に，$k$-way カット[16] を用いてグラフの分割を行う．$k$-way カットとは，$k$ 個の頂点 $s_1, s_2, \ldots, s_k$ をターミナル $T$（$\subseteq V$）と呼ばれる特別な頂点としたときに，そのターミナルを 1 つずつ含むような部分グラフ $G_1, G_2, \ldots, G_k$ に分割する手法である．ただし，このような分割は複数存在するため，式 (1) のようにコスト関数を定義し，これを最小化する分割を求める．なお，$a_{x,y}$ は，頂点 $x$ に対応するベクトルの $y$ 番目の要素（TFIDF 値）を示す．

$$\mathrm{Cost}(G_1, G_2, \ldots, G_k) = \sum_{i=1}^{k} \sum_{\{u \in G_i, v \in V \setminus G_i\}} \frac{\sum_j a_{u,j} a_{v,j}}{\sqrt{\sum_j a_{u,j}^2} \sqrt{\sum_j a_{v,j}^2}}$$

**図 4** に $k$-way カットの例を示す．このグラフでは，3 つのターミナルを与えている．各ターミナル $s_1$，$s_2$，$s_3$ はそれぞれ $G_1$，$G_2$，$G_3$ に含まれ，$\mathrm{Cost}(G_1, G_2, G_3)$ が最小となるようにカットが決定されている．決定されたカットは破線で示されており（図左），これらの辺を除去することで部分グラフが得られる（図右）．

コンテンツ特徴による DBG の分割は，2 クラスのクラスタリング問題と見なせるため，K 平均法など他の方法も適用可能である．本研究で特に 2-way カットを採用した理由は，類似したページをまとめるというクラスタリング的方法ではなく，相対的に類似度の低いページ間の辺を除去することでグラフの分割を行うという 2-way カットの方法が，コンテンツ特徴が雑多であると予想される DBG を大雑把に 2 分割するという目的により適うと考えたことによる．

### 3.4　リンク構造とコンテンツ特徴を用いた効率的ラベリングとスプログ検出

本節では，リンク構造とコンテンツ特徴を複合的に用いたブログとスプログの分離方法についてについて述べる．また，分離したクラスタを利用した効率的なラベル付き訓練データ作成方法，これを利用したスプログ検出についても述べる．

まず，データセット全体からページ間のリンク情報を取得し，リンク構造に基づくグラフを作成する．このグラフから，3.2 節の方法によって DBG クラスタ群を抽出する．続いて，得られた各クラスタに対して，3.3 節で述べた $k$-way カットを適用することで，コンテンツ特徴に基づくクラスタリングを行う．なお，$k$-way カットを適用する際は，ブログとスプログの 2 クラスを想定し，$k = 2$ とする．これら一連の処理によって，リンク構造とコンテンツの両方の観点から特徴的なクラスタ群が抽出できる．より具体的には，コンテンツ特徴に

よって内部的に二分割された DBG クラスタが複数生成される．

次に，各 DBG クラスタ内の代表的なページを見つける．ここでは，コンテンツ特徴に関して他ページと高い類似度を持つページがそのクラスタで代表的なページ（以降「代表ページ」と呼ぶ）であると考える．この代表ページを同定するため，次のように類似度の投票を行う．得られた各 DBG クラスタ内の各ページについて，そのページに最も類似しているページに投票を行う．これをクラスタ内のすべてのページについて繰り返すと，最終的に獲得票数の最も多いページがそのクラスタ内で類似したページを最も多く持つページとなる．本研究では，このページを代表ページと見なす．なお，前節で，各 DBG クラスタはコンテンツ特徴によって 2 つの部分グラフに分割されていると述べた．これら 2 つの部分グラフのうち，以降の処理では，代表ページをメンバとして含む部分グラフだけを訓練事例候補として利用し，代表ページを含まない部分グラフは破棄する．これは，代表ページが属さない部分グラフにはスプログとブログが混在していることが多いため，訓練事例として有用でないことによる．

このようにして得た代表ページのラベル付けを人手で行い，そのクラス（ブログまたはスプログ）を決定する．そして，その代表ページを含む部分グラフ全体のラベルを代表ページのラベルで代える．これにより，本来は部分グラフ内のページすべてをラベル付けする必要があるところを，ただ 1 つのページのラベル付けのみで完了することができる．もちろん，同定した部分グラフ内に実際にはブログとスプログが混在する場合も考えられる．この場合，訓練データの質が低下するため，スプログ検出の精度も低下することになる．この点については，4.1 節の実験によってその影響を検証する．

上述の方法で得た訓練データを用いて分類器を作成し，スプログ検出に用いる．分類器としては，先行研究で優れた結果を残している SVM を使用する．分類器の素性には，訓練データ作成の際に用いた TFIDF を要素とする単語ベクトルを利用する．

## 4. 評価実験

提案手法の有効性を評価するため，実データを用いて 2 つの観点から評価実験を行った．1 つ目の観点は従来手法との比較であり，もう 1 つは，グラフ構造とコンテンツ特徴の複合的な利用の効果である．以下の節で，それぞれについて，実験の詳細と結果および考察を述べる．

実験には，Kolari ら[1] の作成した Splog Blog データセットを使用した．このデータセットは 3,000 のブログからなり，このうち 700 が（スプログではない）ブログ，同じく 700 がスプログとラベル付けされている．本研究では，ラベルが明示的に付与されている 1,400 件のブログを利用した．なお，以下の実験で用いる真のラベルは，このうちスプログ・ブログ各 1 件の計 2 件であり，これらを基に全 1,400 件を評価データとしてラベルの予測を行った．1,400 件のうち 2 件は真のラベルが所与であるため，厳密にはこれらを除いた 1,398 件について評価を行うべきところである．しかし，精度への影響が微少（2/1400=0.0014）であるため，あえて除去していない．

### 4.1 従来手法との比較

#### 4.1.1 実験方法

Splog Blog データセットは HTML で配布されているため，まずデータセットからテキスト情報とリンク先 URL を自動的に抽出した．このとき，コメントに含まれる URL およびトラックバックの URL は，ブログ作成者以外でも挿入可能であり，しばしばブログスパムとして用いられるため，抽出対象から除外した．

続いて，抽出したリンク情報からグラフを構築し，複数の DBG を抽出した．抽出された DBG の中から，特に大きさが中規模程度の（50〜150）のクラスタに注目し，$k$-way カットによる部分グラフ分割および代表ページの同定を行った．これは，含まれるページ数が多い DBG はセンターとなるページが一般的であり，スプログの特徴を表しにくいと考えられること，またページ数が少ない DBG は部分グラフ分割や代表ページの選出の信頼性が低いと考えられるためである．なお，今回は上述のように中規模程度を 50〜150 と恣意的に仮定した．これに対し，実際の運用では，大きさが中央値である DBG を順次（最低 2 個）選択すればよい．

最後に，同定された代表ページのラベル付けを行い，その代表ページを含む部分グラフのラベルとした．なお，代表ページのラベル付けは，スプログ・ブログの部分グラフがそれぞれ少なくとも 1 つ得られるまで行う必要がある．しかしながら，今回の実験では，最小のラベル付けコストで得られるスプログ検出精度を確認するため，2 件の代表ページがそれぞれスプログとブログであるものとした．いい換えると，この 2 件の代表ページにのみラベル付けを行った．なお，本実験で利用したデータセットにはすでにラベルが付与されているため，実際には代表ページに既付与のラベルをそのまま用いてラベル付けを行ったこととした．代表ページ以外のページに関する真のラベル情報は，後述する評価指標値の算出以外にはまったく用いていない．以上の手続きで，ブログ事例とスプログ事例を獲得し，これを基に分類器（SVM）のパラメータを学習した．SVM のカーネルには，経験的に RBF カーネルを用いた．

**表 1**　従来手法と比較したときのスプログ検出の結果
Table 1　Comparative results of splog/blog detection.

| | | 教師あり | | 半教師あり | | 提案手法 |
|---|---|---|---|---|---|---|
| | | 従来手法 1 | 従来手法 2 | 従来手法 3 | 従来手法 4 | |
| スプログ | $R$ | 0.003 | 0.767 | 0.849 | 0.000 | 0.673 |
| | $P$ | 1.000 | 0.832 | 0.495 | 1.000 | 0.732 |
| | $F_1$ | 0.006 | 0.797 | 0.610 | 0.000 | 0.695 |
| ブログ | $R$ | 1.000 | 0.843 | 0.197 | 1.000 | 0.748 |
| | $P$ | 0.500 | 0.786 | 0.669 | 0.500 | 0.705 |
| | $F_1$ | 0.095 | 0.812 | 0.246 | 0.667 | 0.720 |
| 正解率 | | 50.1% | 80.6% | 52.3% | 50.0% | 71.0% |

### 4.1.2 実験結果

実験結果を**表 1** に示す．なお，再現率（$R$），適合率（$P$）は，それぞれ各クラスの真の事例数，予測した事例数に対する正解事例数の比であり，$F_1$ 値は再現率と適合率の調和平均である．また，正解率は，全事例数における正解事例数の比である．従来手法 1〜4 の結果はそれぞれ次のようにして得た．

- 従来手法 1．提案手法で得られた代表ページのみを訓練事例として用い，分類器（SVM）を学習．
- 従来手法 2．提案手法で利用した訓練データと同数の事例に真のラベルを与えて，分類器（SVM）を学習．
- 従来手法 3．コンテンツ特徴に基づき，単純ベイズ法と EM アルゴリズムによって分類器を学習（半教師あり）[8]．
- 従来手法 4．リンク構造に基づき，最小カット法によってデータを分割（半教師あり）[7]．

従来手法 1 は，提案手法と同一の人手で作成したごく少数の訓練事例のみを用い，データの特性，すなわちリンク構造およびコンテンツ特徴を用いていない．従来手法 2 は，提案手法が分類器作成に用いた訓練事例（代表ページがブログの部分グラフとスプログの部分グラフ）と同数の真のラベル付き訓練事例を用いている．よって，従来手法 1 と従来手法 2 は，それぞれ提案手法が持ちうる性能の下限と上限と見なすことができる．一方，従来手法 3 と 4 はそれぞれコンテンツ特徴とリンク構造（グラフ）を用いた代表的な半教師ありの分類法である．前者は少数のラベル付き事例を用いて分類と評価を繰り返すことで漸次的により高精度の分類器を構築し，後者は与えられた 2 つのラベル付き事例を基に 2 値分類を最大フロー問題と見なしてグラフを分割する．2 章で述べたように，提案手法は半教師あり学習手法とも見なせるため，これらの手法と比較することで提案手法の有効性を検証する．

なお，提案手法に関しては，リンク構造から複数の DBG が得られるため，どの DBG を用いて以降の処理を行うかで，分類精度に違いが生じる．そこで，代表ページがスプログの DBG とブログの DBG のすべての組合せに対して分類精度を算出し，表 1 にはその平均のみを示した．従来手法 1 も同様である．また，従来手法 3 と 4 についても，最初に与えるラベル付き事例によって分類精度が異なるため，無作為に選んだそれぞれ 10 組，300 組のブログとスプログについて実験を繰り返し，その平均を表 1 に示した．

従来手法 1 を見ると，スプログ・ブログともに $F_1$ が 0.006，0.095 ときわめて低い．これに対し，提案手法ではそれぞれ 0.695，0.720 の $F_1$ を得た．また，正解率を見ると，従来手法 1 では，約 50%であり，ランダムに分類した場合と変わりがない．すなわち，ごく少数の代表ページのみから分類器を作成したときは学習の効果がまったくないのに対し，提案手法のようにデータの持つ構造・特徴を効果的に利用することで，ラベルが付与されていない事例をも有効な訓練事例として用いることができることが分かる．

さらに，提案手法を従来手法 2 と比較すると，$F_1$，正解率ともおよそ 10%程度の精度低下が見られる．これは，従来手法 2 で利用した訓練事例はすべて真のラベルを持っているのに対し，提案手法ではほとんどの事例のラベルは，代表ページから類推されたものであることによる．当然，類推が誤っている場合もあるため，そのようなノイズがスプログ検出精度の低下に現れている．いい換えると，両者の違いは，人手によるラベル付けのコスト削減と精度の低下のトレードオフであり，どちらが好ましいかはスプログ検出の目的やラベル付けが必要とされる頻度などに依存する．しかしながら，冒頭で述べたスプログの特徴の変動性などを考慮すると，コスト削減の効果は大きいと考えられる．

次に，従来手法 3 および 4 については，半教師ありの手法であるにもかかわらず，同数のラベル付き事例を用いた教師あり手法の従来手法 1 と比較してほとんど精度の向上が得られなかった．特に従来手法 4 では，ほとんどすべてのページがブログと判定されてしまった．分類の結果を調査したところ，この主な原因は，利用したデータセットが比較的小規模なためにリンク構造が疎らすぎたことにあった．従来手法 3 についても，精度の向上はきわめて限定的であった．これは，与えるラベル付き事例が 2 件と極端に少ないため，分類・評価を繰り返しても有効な訓練事例を得ることができなかったためだと考えられる．これらの手法と比較して，提案手法ではコンテンツ特徴とリンク構造を用いることで，ブログ集合とスプログ集合を大まかにグループ化することが可能となり，分類精度の向上につながったものと考えられる．

**表 2**　コンテンツ特徴またはリンク構造のみを使った場合と提案手法との比較
Table 2　Comparison between our approach and those using either contents or hyperlink structure.

| | | コンテンツ | リンク | 提案手法 |
|---|---|---|---|---|
| スプログ | $R$ | 0.510 | 0.937 | 0.673 |
| | $P$ | 0.520 | 0.593 | 0.732 |
| | $F_1$ | 0.520 | 0.722 | 0.695 |
| ブログ | $R$ | 0.530 | 0.354 | 0.748 |
| | $P$ | 0.520 | 0.851 | 0.705 |
| | $F_1$ | 0.520 | 0.481 | 0.720 |
| 正解率 | | 51.9% | 64.1% | 71.0% |

**表 3**　リンク構造によって抽出された DBG 中のブログ・スプログ比
Table 3　Splog/blog ratios within individual DBGs identified by link structure.

| ブログページ数 | スプログ比（$r$） | ブログ比（$1-r$） |
|---|---|---|
| 392 | 0.41 | 0.59 |
| 174 | 0.63 | 0.37 |
| 135 | 0.70 | 0.30 |
| 73 | 0.63 | 0.37 |
| 60 | 0.85 | 0.15 |
| 57 | 0.04 | 0.96 |
| 53 | 0.40 | 0.60 |
| 39 | 0.56 | 0.44 |
| 9 | 0.44 | 0.56 |
| 9 | 0.78 | 0.22 |

### 4.2　リンク構造とコンテンツ特徴の複合的な利用の評価

#### 4.2.1　実 験 方 法

本節では，リンク構造とコンテンツ特徴の両者を利用することで，スプログおよびブログの集合をより正確に同定し，より質の高い訓練データを作成することができたのかをスプログ・ブログ分類の精度から検証する．このために，a) コンテンツ特徴から 2-way カットによって得た部分グラフの代表ページをラベル付けし，このラベルをその部分グラフのすべてのページに付与して分類器を作成した場合，b) リンク構造に基づいて得られた DBG の代表ページをラベル付けし，このラベルをその DBG に含まれるすべてのページに付与して分類器を作成した場合，c) 両方を用いて得られた訓練データを用いて分類を行った場合（提案手法）を比較する．なお，コンテンツ特徴からの部分グラフの作成については 3.3 節で述べた．

この実験により，コンテンツ特徴のみを用いた手法 a)，およびリンク構造のみを用いた手法 b) よりも，提案手法 c) が高い分類精度を示していれば，両者を複合的に利用することでより質の高い訓練データ，すなわち純度の高いスプログおよびブログ集合を同定できたものと考えられる．なお，いずれの場合も，代表ページの選出と代表ページに対するラベル付けは，3.4 節に述べた方法で同様に行う．実験に用いたデータセットは前節と同一である．

#### 4.2.2　実 験 結 果

実験結果を**表 2** に示す．ここで，「コンテンツ」，「リンク」，「提案手法」がそれぞれコンテンツ特徴のみから得た部分グラフ，リンク構造のみから得た DBG，両者を用いた提案手法の結果に対応する．「リンク」，「提案手法」に関しては，複数の DBG が存在するため，前節と同様にすべてのスプログ DBG とブログ DBG の組合せについて実験を繰り返し，平均値を報告している．

表 2 の「コンテンツ」を見ると，スプログ・ブログともほとんど分類の効果が現れていないことが分かる．この結果は，コンテンツ特徴のみによってデータ全体を二分割しても，スプログとブログをうまく分離することは難しいことを意味する．この原因として，スプログは注目キーワードや他のウェブページのコンテンツを利用して作られることが多いことに起因すると考えられる．

次に「リンク」の結果を見ると，スプログに関する再現率（$R$）が 0.937 と大きく向上している反面，ブログに関する再現率は低下している．すなわち，リンク構造のみを用いた手法では，多くのページをスプログと誤判定してしまっていることが分かる．リンク構造を用いた手法について，さらに詳細に結果を分析するため，抽出された DBG に含まれていた真のスプログとブログの比率を**表 3** に示す．ここで，「ブログページ数」は抽出された DBG 内に含まれるブログページ数であり，「スプログ比」，「ブログ比」は DBG 内のスプログとブログの比率を示す．これらの比率は，Splog Blog データセットに人手で付与された真のラベルに基づいて算出した．この表から，抽出された大部分の DBG は，スプログ・ブログのどちらかのページを多く含む一種のコミュニティを構成していることが分かる．特に，一部の DBG に関しては，スプログ比が 0.96 やブログ比が 0.85 と非常に純度の高いコミュニティが抽出されている．このような分布から，一部の高い純度の DBG を訓練データとして選択した場合にはそれなりに高い精度でスプログ検出が行えるのに対し，そうでない場合には検出精度が低下すると考えられる．この問題点に関しては，本節で再度言及する．

続いて，表 2 の「提案手法」と「リンク」を $F_1$ によって比較すると，提案手法ではブログ検出の精度が 0.481 から 0.720 へと大きく向上し，全体の正解率も「リンク」の 64.1%か

**表 4**　リンク構造のみを使った場合と提案手法の正解率の比較
Table 4　Comparison between our approach and link-based approach in terms of stability.

| | リンク | 提案手法 |
|---|---|---|
| 平均 | 64.1% | 71.0% |
| 標準偏差 | 4.55 | 1.78 |

ら 71.0%へと 10%程度向上している．一方，再現率に注目すると，ブログについては 0.354 から 0.748 へ倍増するとともに，スプログについては 0.937 から 0.673 へと低下している．どちらが好ましい挙動かはスプログ検出の目的によるものの，ブログの再現率があまりにも低い場合，有用なブログがスプログとして誤検出・排除されるという弊害が頻繁に生じることになる．

最後に，前述の問題点について考察する．リンク構造のみを用いた場合，スプログ検出の精度が，訓練データとして利用する DBG に強く依存する可能性があることを述べた．この点に関して，リンク構造のみを用いた場合と提案手法を比較した．**表 4** に，両者の正解率の平均（表 2 の「正解率」と同一），分散，標準偏差を示す．この結果が示すように，「リンク」の正解率の標準偏差が 4.55 なのに対し，提案手法では 1.78 と大幅に減少している．この結果から，提案手法では，リンク構造に加えてコンテンツ特徴を用いることで，任意の DBG 中でさらにスプログあるいはブログの純度の高い部分グラフを見つけることができていることが分かる．その結果，訓練データとして利用する DBG の違いによらず，比較的安定した性能が得られているものと考えられる．

## 5. むすび

本研究では，プログの 2 つの特性，すなわちリンク構造とコンテンツ特徴に着目し，これらの特性を複合的に用いることで，スプログ・ブログ分類器作成のための訓練データ構築に要するコストを大幅に削減する手法を提案した．より具体的には，ブログ間のリンク関係から生成されるグラフから DBG を抽出し，さらに DBG をコンテンツ特徴によって二分割することで，スプログまたはブログが多く含まれるページ集合を獲得した．そして，そのようなページ集合から，コンテンツ特徴の類似度による投票から代表的なページを同定し，この代表ページのみに人手でブログかスプログかの判定を行った（ラベルを与えた）．そして，ページ集合の他のページへも代表ページと同じラベルを与え，分類器の訓練データとした．

実データを用いて提案手法の有効性を評価したところ，人手で付与したラベルのみを用いて分類器を学習した場合と比較し，正解率は 4 割以上増加した．また，提案手法と同数の（真のラベルを持つ）訓練事例を与えた場合と比較しても，正解率は 1 割程度の減少にとどまった．いい換えると，わずか数ページだけ人手でラベル付けを行えばよいというコスト削減の効果に比べ，精度への悪影響は限定的であった．

さらに，リンク構造とコンテンツ特徴を複合的に用いたことによる効果を検証するため，それぞれの特性のみを用いて，提案手法と同様の手順で訓練データの構築を行い，スプログ検出の実験を行った．その結果，コンテンツ特徴のみではスプログ検出にはまったく効果がなく，リンク構造のみでは訓練データとして用いる DBG によってスプログ検出の精度が大きく変化した．提案手法では，リンク構造を用いて DBG を抽出した後にコンテンツ特徴によって部分グラフを得るという方法によって，より高精度かつ安定した結果が得られていた．今後の課題として，より大規模なデータを用いた評価を計画している．また，本研究の主眼はコンテンツとリンクの複合的利用の効果の検証にあったものの，今後は，DBG や 2-way カット以外の代替アルゴリズムも考慮したスプログ・ブログ集合の最適な抽出方法の検討が必要である．

## 参考文献


1) Kolari, P., Finin, T. and Joshi, A.: SVMs for the blogosphere: Blog identification and splog detection, *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs* (2006).
2) Lin, Y.-R., Sundaram, H., Chi, Y., Tatemura, J. and Tseng, B.L.: Splog detection using self-similarity analysis on blog temporal dynamics, *Proc. 3rd International Workshop on Adversarial Information Retrieval on the Web*, pp.1–8 (2007).
3) Kolari, P., Java, A. and Finin, T.: Characterizing the splogosphere, *Proc. 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics* (2006).
4) Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web search engine, *Proc. 7th International World Wide Web Conference* (1998).
5) Ishida, K.: Extracting spam blogs with co-citation clusters, *Proc. 17th International Conference on World Wide Web*, pp.1043–1044 (2008).
6) Ikeda, D., Takamura, H. and Okumura, M.: Semi-supervised learning for blog classification, *Proc. 23rd National Conference on Artificial Intelligence*, pp.1156–1161 (2008).
7) Zhu, X. and Goldberg, A.B.: *Introduction to semi-supervised learning*, Morgan & Claypool Publishers (2009).
8) Nigam, K., McCallum, A. and Mitchell, T.: *Semi-supervised learning*, chapterSemi-supervised Text Classification Using EM, MIT Press (2006).

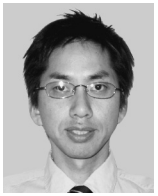
9) Tong, S. and Chang, E.: Support vector machine active learning for image retrieval, *Proc. 9th ACM International Conference on Multimedia*, pp.107–118 (2001).
10) Ounis, I., de Rijke, M., Macdonald, C., Mishne, G. and Soboroff, I.: Overview of the TREC-2006 blog track, *Proc. 15th Text Retrieval Conference* (2006).
11) Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J.: Graph structure in the Web, *Proc. 9th International World Wide Web Conference on Computer Networks*, pp.309–320 (2000).
12) Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A.: Trawling the Web for emerging cyber-communities, *Proc. 8th International Conference on World Wide Web*, pp.1481–1493 (1999).
13) Reddy, P.K. and Kitsuregawa, M.: An approach to relate the Web communities through bipartite graphs, *Proc. 2nd International Conference on Web Information Systems Engineering*, pp.301–310 (2001).
14) Flake, G.W., Lawrence, S., Giles, C.L. and Coetzee, F.M.: Self-organization and identification of Web communities, *Computer*, Vol.35, No.3, pp.66–71 (2002).
15) Salton, G. and McGill, M.J.: *Introduction to modern information retrieval*, McGraw-Hill, Inc. (1983).
16) Dahlhaus, E., Johnson, D.S., Papadimitriou, C.H., Seymour, P.D. and Yannakakis, M.: The complexity of multiterminal cuts, *SIAM Journal on Computing*, Vol.23, No.4, pp.864–894 (1994).

(平成 21 年 9 月 18 日受付)
(平成 21 年 12 月 25 日採録)

(担当編集委員　馬 強)

**吉川　幹人**
平成 21 年神戸大学大学院工学研究科情報知能学専攻博士前期課程入学.

**佐藤　翔平**
平成 21 年神戸大学大学院工学研究科情報知能学専攻博士前期課程修了. 同年 NTT ドコモ株式会社入社.



**関　　和広**（正会員）
平成 14 年図書館情報大学情報メディア研究科修士課程修了. 平成 18 年インディアナ大学図書館情報学研究科博士課程修了. 同年より神戸大学助手（現，助教）. 情報検索，自然言語処理，機械学習の研究に従事. Ph.D. 電子情報通信学会，自然言語処理学会，ACM SIGIR 各会員.

**上原　邦昭**（正会員）
昭和 53 年大阪大学基礎工学部情報工学科卒業. 昭和 58 年同大学院博士後期課程単位取得退学. 同産業科学研究所助手，講師，神戸大学工学部情報知能工学科助教授，同都市安全研究センター教授を経て，現在，同大学院工学研究科教授. 工学博士. 人工知能，特に機械学習，マルチメディア処理の研究に従事. 人工知能学会，電子情報通信学会，計量国語学会，日本ソフトウェア科学会，AAAI 各会員.