リンク構造とコンテンツを相補的に用いた極少訓練事例からのスプログ発見

佐藤　翔平† 　関　和広† 　上原　邦昭†

† 神戸大学　〒 657-8501　神戸市灘区六甲台町 1-1
E-mail: †sato@ai.cs.kobe-u.ac.jp, ††seki@cs.kobe-u.ac.jp, †††uehara@kobe-u.ac.jp

あらまし　インターネットの急激な普及と検索技術の向上によって，ウェブページの商業利用が盛んになり，ブログサイトにおいても高い検索順位を得ることで目的の商業用ページに誘導することだけを目的としたスパムブログ（スプログ）が大量に作られ，検索結果に有害な影響を及ぼしている．このスパムブログの除去のために，従来の手法では自動的に作成されるスプログに関してコンテンツに着目した分類手法が数多く提案されてきた．しかし，このような分類を行うためにはラベル付けされた訓練事例が必要となり，高いコストが必要となる．そこで本論文では訓練事例の候補となるような類似的なページ集合を見つけ，その中の一部のページを評価することで訓練事例作成のコストを下げた手法を提案する．そのためにまずブログサイトのリンク構造に着目し，DBG アルゴリズムによりリンク先が類似的なページ集合を取得した．また，ブログサイトとスプログのテキスト特徴は異なるという仮定からテキスト特徴について類似的なページ集合を発見し，リンク構造と相補的に用いることで類似的なページ集合を発見する．その中のとりわけ特徴的なごく一部のブログのみにラベルを与えることで効率的にスプログを発見する手法を提案する．
キーワード　スプログ ブログ　機械学習　クラスタリング

A splog detection framework exploiting link structure and contents with few labeled data

Shohei SATO†, Kazuhiro SEKI†, and Kuniaki UEHARA†

† Kobe University　Rokoudai-cho 1–1, Nada-ku, Kobe, 657–8501 Japan
E-mail: †sato@ai.cs.kobe-u.ac.jp, ††seki@cs.kobe-u.ac.jp, †††uehara@kobe-u.ac.jp

Abstract　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 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.
Key words　splog, blog, machine learning, clustering

1. ま え が き

現在インターネットは急速な発展を遂げており，生み出されるコンテンツの中には個人の意見を多く含むものも多数存在する．そのような意見を含むウェブページの中でも一般的なものがウェブログ（ブログ）である．

ブログはニュース，製品へのレビューや日記など個人の意見が多く含まれるため，近年注目されているコンテンツ資源である．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リンク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作成されるようになった．

これらスプログは検索結果上位に表示されるように検索エンジン最適化の手法を用いたり，話題の単語を無作為に用いたりと一般的に検索結果に悪影響を及ぼす．

このような問題を引き起こすため，スプログの発見，検出が既存の研究において試みられてきた．例えば Kolari ら [7] 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 Yu ら [11] 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発見の手法を提案している．これらの手法にはいずれも機械学習が用いられているため，事前に訓練データを必要とする．しかし，必要となるクラス付けされた訓練事例を人手を使って得るには多大なコストが必要となる．また，スプログのコンテンツ特徴は日々変化していくことが考えられる．これは検索上位となるようなトピックはその時々で変わっていくためである．このようにトピックが変化した場合に過去に作った訓練事例がスプログの特徴を必ずしも表すものではなくなると考えられる．こうした場合に即座に新しい訓練事例を作成することは前述のとおり容易ではない．

そこで，このコストを減らすことが本研究の目的である．本論文で提案するモデルはスプログとブログの間の違いに着目し，類似的な構造を持つページ集合をクラスタリング手法によって抽出する．さらに，得たクラスタの代表的なページを見つけ，そのページにクラスタに含まれるページ全体の評価を一任する．このようにすることで訓練データを作成する際にごく一部のページを評価するだけで済むため，人手のコストは大幅に軽減される．

本論文ではまずスプログの定義と着目した構造についての詳細を述べる．次に，どのようにして特徴的な構造を用いて類似的なページ集合を得るかについての説明を行う．そして提案手法の有効性を調べるため，ページ集合の評価と実際に評価データを用いてスプログの検出を行う．

2. Splog の特徴に基づくクラスタの仮定

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 [5] はスプログ研究において定義付けを行っており，そこではスプログを作成するための目的として 2 つの動機を挙げ，このような動機において作成されたブログページをスプログとしている．

（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） 提携するサイトの検索エンジンランキングを押し上げる

スプログの作成にはこのような動機が背景にあるため，スプログの記事は製品やサービスに関する説明のないアフィリエイト広告へのリンクだけであったり，他のブログの複数記事を断片的に貼り付けたページになることが多い．ただし，ある種の商品へのリンクや商業目的のサイトへのリンクが多いからといっただけでスプログとみなされる訳では無い．ブロガーの体験に関連する広告や実際の店舗のスタッフなどによる商品入荷などの情報を掲載しているようなページはブログとみなされる．あくまで目的のサイトへのリンクだけを目的とし，自ら情報を発信しないページがスプログとみなされる．

図 1，図 2 にスプログの例を示す．図 1 では単純にリンク先だけの記事しかなく，見た目からも容易にスプログと判断できる例である．このようなページはユーザーにリンク先に誘導することだけを考えている．

図 2 は図 1 に比べてやや巧妙なスプログである．一見普通のブログサイトのように見せかけられており一目では判断できないが，実際にはスプログでありリンクをたどると商業目的のページが多い．また良く見ると本文も不自然であることが分かる．

図 1　スプログの例 1

図 2　スプログの例 2

2.1 スプログ検出における関連研究

スプログ発見の従来研究について説明を行う．Kolari ら [7] はスプログ空間に対して調査を行い，スプログ抽出のための手法として SVM (support vector machine) を用いて F 値 0.87 のスプログ発見に成功している．この研究ではブログ記事が持つ URL やアンカーテキストを新しく分類器の素性として提案しその効果についても論じている．またスプログの単語特徴にも着目しこれらの素性と提案素性とを組み合わせて使用してい

る．また，Kolari らとは別のグループである Yu ら [11] は投稿時間の分布，内容・リンク先の情報を基にし，ブログ記事の自己相似性を用いてスプログ発見を行う手法を提案し，AUC0.9 以上の性能を出している．

これら従来手法で提案されるスプログ分類手法は静的な訓練データを元にした分類手法であり, その対象とされているコンテンツ分布に特化されているため，異なるコンテンツ分布に最適な分類手法とは言えない．しかし，スプログは日々大量に投稿されている上にスプログへの検索を誘導するための注目されるキーワードは時間が経つごとに変化していくと考えられる．このように新たなスプログ空間が作成されうる場合，一つの解決案として再び目的にあった分布に特化したデータセットを用意すればよいと考えられる．しかし，既存の手法では訓練データセットを用意するためには多大なコストがかかるため，簡単にデータセットを作成するわけにはいかない．

このため，従来研究などのように大きなコストをかけずに訓練データを作成し，スプログの発見ができれば有用であると言える．

2.2 スプログの構造的な特徴

前節ではスプログの概要について述べ，いくつかの特徴をあげた．これらの特徴からスプログの特徴に関していくつかの所見が得られる．一つ目は，スプログとブログはそれぞれリンク構造に特徴があるのではないかというかという推察である．

スプログの目的のひとつは特定のアフィリエイト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ある一定のコミュニティであり，類似的なページ集合である可能性がある．またブログは類似的なブログ同士でリンク構造をなしており，このようなコミュニティではスプログは排除されやすいと考えられる．このことからスプログ・ブログともにリンク構造としてある程度の特徴を持ち，それぞれに独自のコミュニティを形成していることが推測される．

二つ目に，従来の研究からコンテンツの類似度に注目することでスプログとブログを分けることができる可能性がある．

これらの所見から，リンク構造とコンテンツ構造に着目することでスプログとブログのページ群が抽出できれば，目標となる小さなコストの訓練データ作成に役立つのではないかと考えた．この仮定が正しければ，リンク構造とコンテンツ構造の二つの尺度でページを分割，抽出することでスプログかブログのどちらかの特徴に偏ったページ集合が得られることになる．そしてそれらの特徴的なページ集合内の代表的なページを見つけることが出来れば，そのページの評価を行うだけでページ集合全体の評価が行える．こうすることで人手で全てのページを見る必要は無くなるためコストを大幅に削減できる．

次章ではこのスプログの構造に着眼点を得た極小コストでの訓練データ作成手法について具体的な説明を行う．

3. 提 案 手 法

前章ではスプログの概要とその発見における従来研究について述べた．従来研究ではスプログ発見のための分類を行う際に訓練データを収集する必要があり，そのために大きなコストがかかることを示した．

そこで我々はスプログのリンク構造とコンテンツ構造に着目し，それぞれの観点からクラスタリングを行うことを提案する．そして，それぞれの尺度で得られたクラスタを組み合わせることで訓練データ候補となるクラスタを作成し，その中のごく．一部のデータを評価することで訓練データを作成する．

以下でそれぞれの構造に関して適応するクラスタリング手法について説明を行う．また，これらのクラスタリング手法を相補的に用いるスプログ分類の手法についても説明する．

3.1 リンク構造によるクラスタリング

ブログとスプログのリンク先の集合は異なる可能性があることについて前述した．スプログは特定の商業サイトにリンクを行っている可能性が高く，またこれらのサイトはお互いにリンクされることで一種のコミュニティを形作っているのではないかという推察である．このようなリンク構造に着目すると，ある一定の web コミュニティを参照しているページ集団を抽出することでこのような構造を抽出できる．そこで，リンク構造に関するクラスタリングに web コミュニティ抽出手法を適用する．以下ではコミュニティ抽出の具体的な手法に関して説明する．

3.1.1 ウェブコミュニティ

ウェブコミュニティは,「共通のトピックを持ったウェブページの集合」と定義されるページ集合のことである．ウェブコミュニティ抽出はウェブのリンク構造から特徴的な部分構造を抽出することでウェブコミュニティを抽出し，WWW を意味的に分類・整理することを目的としている．この目的は本研究の目的であるリンク構造から特徴的なクラスを発見するという考えに合致している．そのため，ウェブコミュニティ抽出の手法をリンク構造によるクラスタリングに対して応用を行う．

ウェブページを頂点とし，リンクを辺とするとウェブ空間を巨大な有向グラフとしてとらえることができる [1]．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ウェブグラフから，ウェブコミュニティに対し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．Kumar らの提案した trawling [10] では，ウェブコミュニティを,「十分に大きく，十分に密な 2 部グラフ」として定義し，データセットから全ての完全 2 部グラフを抽出する手法を提案している．このような２部グラフを発見するための手法として以下で説明する Dense Bipartite Graph がある．

3.1.2 Dense Bipartite Graph

DBG は密な二部グラフをウェブコミュニティとした手法で，Reddy らによって提案された．Reddy らの手法で言われている密な二部グラフとは，ファンはセンターに対して閾値 p 以上のリンクを持ち，センターはファンから閾値 q 以上のリンクを持つ二部グラフ（DBG: Dense Bipartite Graph）を意味する．DBG を抽出するアルゴリズムは，シードページからリンクをたどり DBG の候補を抽出し，そこから DBG を抽出することである．具体的には，まずシードページ s を選んで集合 $S = s$, 集合 $T = \phi$ とし，S からリンクされているページを T に加えて再び T をリンクしているページを S に加えることを繰り返

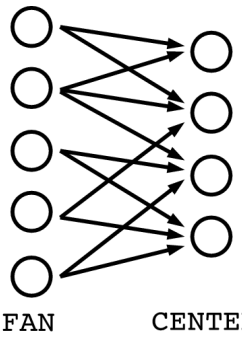

図 3　DBG の例 (p=2, q=2)

し行い，密な二部グラフの候補とする．そして $s \in S$ に対し，T に対する s のリンクの数が一定値以下なら S から削除し，$t \in T$ に対しても S からのリンクの数が一定値以下なら T から削除，ということを繰り返す．Reddy らは DBG によるウェブコミュニティの抽象化をウェブコミュニティ同士に適用することにより，関連する web コミュニティを抽出している．図 3 に DBG コミュニティの例を示す．

また，本研究での DBG の具体的な抽出手順を以下に示す．

（1） ブログ集合をウェブページ集合 F とする．

（2） Fan の持つアウトリンク数の閾値 p と，Center の持つインリンク数の閾値 q を与える．

（3） STEP 1 により得られたウェブページ集合 F を Fan とし，Fan からリンクされているウェブページ集合を Center とする．

（4） Fan と Center を構成するウェブページ数が共に収束するまで，以下の手順 (a)，(b) を繰り返し，DBG の定義を満たすウェブページのみを抽出する．

（a） **C**enter に含まれるウェブページに対して，p 未満のリンクしか持たない Fan 中のウェブページを削除する．

（b） **F**an に含まれるウェブページから，q 未満のリンクしか受けていない Center 中のウェブページを削除する．

（5） 最終的に残された Fan と Center をウェブコミュニティとして出力する．

本研究ではこの DBG 抽出手法を用いてリンク構造に対してクラスタリングを行う．

3.2　コンテンツ構造によるクラスタリング

ここではコンテンツ構造に着目したクラスタリング手法について説明する．前節でスプログとブログのコンテンツ構造に違いがあることについて述べた．そこで特徴的な単語を素性として選択し，TFIDF を計算した．そして各ページ同士の間のコサイン類似度を計算した，ここで各ページ間のコサイン類似度はページ同士の類似性を表す．

3.2.1　グラフ分割

ページ間の類似度はコンテンツに関するページ間の距離とみなすことが出来る．類似度が高ければコンテンツに関する距離が近く，低ければ距離が遠いとみなすと，距離が近いページ同士の集合を求めることでコンテンツとして類似的な集合を作ることが出来る．

そこでページ間の距離を元にグラフ分割の手法を用いて距離の遠いページ同士の間でグラフを分割し，距離の近いページの集合を得る．このような考えからスプログのコンテンツ構造に大してグラフ分割の手法を用いてクラスタリングを行う．

3.2.2　コンテンツ構造を対象としたグラフ構造

データの分布構造を表現するために，データセットにおける各データを頂点とし，その類似度を辺の重みとするグラフを構成する．それは重み付き無向グラフ $G = (V, E)$ で表現される．ここで，V は頂点集合，E は辺集合である．グラフ G は，どの辺も異なる頂点を結ぶという制約を与えており，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のない単純グラフとなる．また，各頂点 u, v をつなぐ辺 $e(\in E) = \{u, v\}$ には，後述する重み関数によって重み $weight(u, v)$ が与えられるとする．各辺の重みは，両端の頂点間の類似度を表す．本研究では類似度の計算手法として，コサイン類似度を使用する．

3.2.3　k-way カット

k-way カット [3] とは，あるいくつかの頂点をターミナル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に分割するカット手法である．ここで，複数の部分集合に分割する場合に用いるため，コスト関数を式 1 のように定義しておく．式 1 において，$A_{u,i}$ は，頂点 u に対応するデータの，i 番目の特徴を示す．

$$Cost(G_1, G_2, ..., G_k) = \sum_{i=1}^{k} \sum_{\{u \in G_i, v \in V \setminus G_i\}} \frac{\sum_i A_{u,i} A_{u,i}}{\sqrt{\sum_i A_{u,i}^2} \sqrt{\sum_i A_{v,i}^2}} \quad (1)$$

k-way カットを求めるには，まずターミナル $\mathrm{T}(\subseteqq V) = \{s_1, s_2, ..., s_k\}$ と呼ばれる特殊な頂点を決定する．各ターミナル s_i を含む部分グラフを G_i すると，$Cost(G_1, G_2, ..., G_k)$ が最小となるようなカットが k-way カットである．

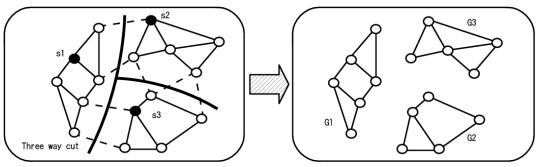

図 4　3-way カットの例

図 4 に実際に k-way カットを算出した例を示す．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 をそれぞれ 1 つずつ含み，$Cost(G_1, G_2, G_3)$ が最小となるカットを決定する．決定されたカットは図 4 左に破線で示したもので，このカットにあたる辺を除去することで部分グラフに分割する (図 4 右)．

3.3 クラスタを利用した極小訓練事例からのスプログ発見手法

本節ではリンク構造に関するクラスタとコンテンツ構造に関するクラスタを利用して，どのようにして訓練データを作成するかについて述べる．また，それと共にクラスタを利用して作成した訓練データを用いてのスプログ発見手法について詳細を述べる．

3.3.1 クラスタを利用した訓練データ候補の選出

まず始めにデータセット全体に対してページ間のリンク情報を取得しグラフ構造を作成する．このグラフ構造に関して DBG を用いてウェブコミュニティの発見を行いクラスタを抽出する．

さらにここで得られたクラスタに対してコンテンツ構造に関するクラスタリングを行うことでリンク情報を相補的に用い，訓練データの候補となる特徴的なクラスタを抽出する．

コンテンツ構造に関してクラスタリングを行うため TFIDF の計算を行い，各ページに関して単語ベクトルを得る．そしてこの特徴量をもとに各ページ間のコサイン類似度を計算する．コサイン類似度は各ページ間のコンテンツに対しての距離量を表し，全体でコサイン類似度を距離としたグラフ構造をなす．このグラフ構造に対して k-way カットを利用したクラスタリングを行う．今回得たいクラスは Blog，Splog の 2 クラスであるため $k = 2$ とする，実際のクラスタリングにはグラフ分割ツール Metis [4] を使った．この操作により，DBG クラスタを二分割したクラスタが得られる．

ここで各 DBG クラスタごとに得られた二つのコンテンツに関するクラスタは後述する代表ページを持つ方だけを選出し，代表ページを含まないものを破棄する．こうして最終的な訓練データ候補が得られる．

3.3.2 代表ページの選出

ページ間の類似度を元にクラスタ内の代表的なページを見つける．より多く，他ページとの間で高い類似度を持つページがそのページ集合で代表的なページと考えることができる．そのため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と類似度の高いいくつかのページを選択し，それらに投票を行う．これをクラスタ内の全てのページについて行うと，投票数の最も多いページが類似度の近いページを最も多く持つページとなる．そこで，このページはクラスタ内のページ集合をコンテンツ特徴の面で最も代表しているページだと考えることができる．このためこれを「代表ページ」と呼ぶことにし，代表ページを含むクラスタだけを用いることにする．こうすることによって，よりコンテンツ特徴が近いデータ集合が得られることになり，訓練データの質が高まると考えられる．

3.3.3 代表ページの評価

最終的に代表ページのラベル付けを行い，そのクラスを決定する．代表ページのラベルをそれを含むページ集合全体のラベルとする．これにより本来はクラスタ内のページ全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だけで行える．このように本手法では従来非常に大きかったラベル付けのためのコストを大幅に削減することができる．

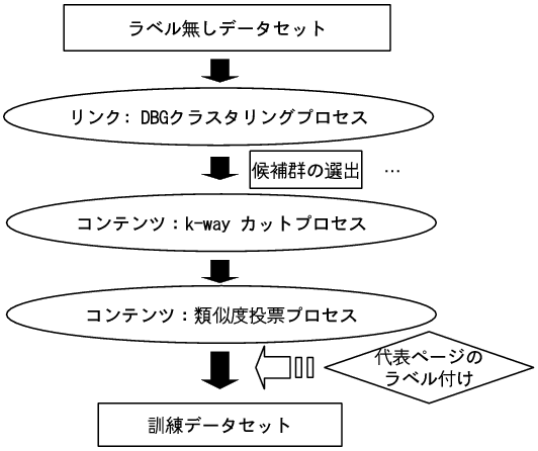

図 5　提案手法のまとめ

3.3.4 スプログ分類

最後に上記の方法で得た訓練データを用いて分類器を作成し，データセット全体を分類する．分類器として従来手法で優れた結果を残している SVM を使用する．分類のための特徴として訓練データ作成の際に計算した単語ベクトルを利用する．

図 5 にシステムの全体的な流れを示す．システムの概要として，ラベルの付いていないデータセットに対し DBG によりクラスタリングを行い，その中からいくつかの候補を選択する．これが DBG クラスタリングプロセス である．次にこれらの候補に対してコンテンツ構造に基づいたクラスタリングを行う，これが k-way カットプロセス である．そして最後の類似度投票プロセス で上記までのプロセスで得られたクラスタから代表ページを選出しラベル付けを行う，これらの処理によってスプログとブログの二つのクラスタを選び，訓練データとしてデータセットに対して分類を行う．

4. 評 価 実 験

4.1 実験 1 - 従来手法との比較実験 -

ここでは訓練データ作成のプロセスの妥当性と実際のスプログの検出精度を検証するため従来手法との比較実験を行う．実験には Kolari ら [7] の作成したデータセットを使用した．データセットには同数のブログとスプログのラベルの付いたブログページそれぞれ 700 ページからなっている．本節では，このデータセットを評価データとして評価実験を行う．

4.1.1 実 験 方 法

評価データは html で記述されているため，データセットから html 解析を行いテキスト情報とリンク先 URL を抜き出した．この時，コメントやトラックバックの URL はブログ作成者以外のユーザでも作れるため除外した．

抜き出したリンク情報から DBG アルゴリズムを用いてクラスタリングを行い，得られたクラスタの中からページ数 50〜150 のクラスタを選び出す．これは含まれるページ数があまりに多いクラスタはセンターとなるページ集合が一般的になりすぎるため，スプログの特徴を表さないと考えたため除外した．また，逆にページ数が少ないクラスタに関しては学習データとするには不十分だと考え除外した．このとき得られたクラスタ

表 1　従来手法との比較

	従来手法 (代表ページ)	従来手法 (全ラベル付き)	提案手法
スプログ:再現率	0.003	0.767	0.663
スプログ:精度	1.000	0.832	0.753
スプログ:F 値	0.006	0.797	0.705
ブログ: 再現率	1.000	0.843	0.782
ブログ: 精度	0.500	0.786	0.699
ブログ: F 値	0.095	0.812	0.738
正解率	50.1%	80.6%	72.3%

に含まれるページ集合に関してベクトル空間モデル [9] を用いて，各ページ同士のコサイン類似度を計算する．また，それと同時に類似度を元にページ同士で高い類似度をもつページに関して投票を行い，最も投票数の多かったページをクラスタ内の代表ページとする．

次にコサイン類似度を距離とみなして k-way カットを用いて二つのクラスタに分割し，代表ページが含まれるクラスタを類似性の高いページ集合として使用する．代表ページが含まれなかったクラスタは使用しない．

最後に得られたクラスタ内の代表的なページをそれぞれブログかスプログのラベル付けを行う．代表ページへのラベルが代表ページを含むクラスタ全てのページのラベルとなる．このようにしてブログ，スプログのそれぞれのページ集合が得られたところでこれを訓練データ集合とする．

この訓練データを用いて分類器の作成を行う．素性として TFIDF を用い訓練を行う．分類器として今回は従来研究 [7] で優れた結果を残している Support Vector Machine (SVM) を使用した．

また Kolari らの単語を素性として SVM で学習を行う手法を比較対象として選択した．提案手法との性能評価として，提案手法で人が評価したデータと同じ量のラベル付きデータを用いた場合の従来手法の分類結果との比較を行う．次に提案手法で評価データからラベル情報をクラスタに拡充した訓練データと同量のラベル付きデータを使って学習した場合の結果と，本手法における分類結果の性能の比較を行い，その性能について比較と評価を行う．

4.1.2　実 験 結 果

実験結果を表 1 に示す．表左の従来手法 (代表ページ) が提案手法で人が評価したデータを訓練データとした場合の結果，従来手法 (全ラベル付き) が提案手法で抽出されたクラスタを使って作られた最終的な訓練データと同量のラベル付きデータを用いて学習した結果である．なお，従来手法 (全ラベル付き) に関しては訓練データをランダムに抽出し，分類した結果の平均を取った．表 1 には分類されたブログとスプログの再現率，適合率，F 値 とともに，分類結果がデータセット全体のデータに対してどの程度正解したかを示す正解率を示している．ここでは従来手法の訓練データセットとして提案手法の評価フェーズで評価された代表ページを用い，同じラベル付きデータを用いたときの二つの手法の性能を比較した．

まず表 1 より従来手法 (代表ページ) を見るとスプログ，ブログともに F 値が 0.006，0.095 と極めて低い事が見て取れる．たいして提案手法ではスプログ，ブログの F 値はそれぞれ 0.705，0.738 と共に従来手法 (代表ページ) に比べて高い．ここからラベルつきデータがわずかしかない場合には従来手法では分類精度が極めて低くなることが見て取れる．それに対して本提案手法ではラベル付きデータが非常にわずかしかないにもかかわらず，F 値に関しては従来手法を大きく上回り，正解率も 22 %向上しており，大幅に分類精度が向上している．

これは代表ページに付加したラベルを抽出クラスタに還元することが訓練データの拡充に有用であり，少ないデータの評価を行うだけである程度のスプログ分類が行えていることを示すといえる．対して従来手法ではラベルデータが極端に少ない場合では訓練データが少なすぎるため分類が十分に行えず，分類精度を向上させるにはコストをかけて訓練データを増やす必要が生じる．

このように，ごく一部のデータだけを評価したにすぎないにも関わらず，従来手法と比較して高い分類精度を得ていることから本手法の考え方が正しく，リンク構造とコンテンツ構造に基づいたクラスタリングによって訓練データの作成コストを抑えたスプログ分類が行えることが示された．

次に従来手法 (全ラベル付き) と提案手法を比較すると従来手法 (全ラベル付き) の F 値はそれぞれ約 0.8 程度なのに対し，提案手法では約 0.7 程度となっており，従来手法の方が優れている．この結果から同量の訓練データがある場合は従来手法の分類精度が上回り，逆にわずかのデータから分類を行った場合は提案手法が優れていると言える．提案手法は本研究の目的である人による訓練データの作成のためのコストを下げることには成功しているが，そのためにやや分類精度が犠牲になっている．

4.2　実験 2 -グラフ構造とコンテンツ構造の相補的な利用の評価-

次に提案手法の仮定の正しさの検証と，実際の動作の検証，そしてクラスタの相補的な利用の結果について検証する．実験のデータセットは実験 1 と同じデータセットを利用する．

4.2.1　実 験 方 法

ここでは提案手法について 3 つの観点から検証を行う．まず一つめは DBG によって得られるクラスタがどのような集合かの検証を行う．このために実際に DBG フェーズで出力されるクラスタを分析し，仮定が正しかったかを検証する．

次にリンク構造とコンテンツ構造の相補的な使用が有効に機能しているかを検証する．このために，データ全体を DBG でクラスタリングして得られたクラスタを訓練データとして分類を行った結果とグラフ分割の手法を用いて得られたクラスタ，そして最後に両方を用いて得られた訓練データを使った場合の分類結果，これら 3 つを比較することで性能の検証と考察を行う．

4.2.2　実 験 結 果

まずはじめにリンク構造とコンテンツ構造，そしてその両方を相補的に用いて訓練データを作成し，分類を行ったそれぞれの場合の結果を表 2 に示す．左からコンテンツ構造をもとに得

表 2　各クラスタを学習データとしたときの結果

	コンテンツ	リンク	コンテンツ＆リンク
スプログ:再現率	0.51	0.937	0.673
スプログ:精度	0.52	0.593	0.732
スプログ:F 値	0.52	0.722	0.695
ブログ: 再現率	0.53	0.354	0.748
ブログ: 精度	0.52	0.851	0.705
ブログ: F 値	0.52	0.481	0.720
再現率	51.91%	64.06%	71.02%

られたクラスタを訓練データとして使用したもの，リンク構造だけを訓練データとしたもの，そして二つを使用した場合の結果を示している．リンク構造に関してクラスタリングを行った場合，複数のクラスタが得られるため，全ての組み合わせに対して結果を求め，その平均を示している．

表 2 からデータ全体に対してコンテンツ構造からクラスタリングを行った場合，分類結果が悪くクラスタリングが適切に行えていないことを表している．これはデータセット全体に対してコンテンツのコサイン類似度を距離量としてクラスタリングを行うと類似的なページが多く，誤まってクラスタリングされてしまうと考えられる．また，データセット全体でクラスタリングを行った場合，訓練データとテストデータが同じになってしまうため分類精度を向上させることが難しいと言える．このことからデータセット全体に対してコンテンツをもとにクラスタリングを行っても有用な訓練データを選別することが難しいことが言える．次にリンク構造から得られたクラスタを訓練データとした場合を見るとスプログの F 値が 0.722 とコンテンツ構造をもとにした場合の F 値 0.52 と比べてかなり向上しておりスプログの分類に関しては優れていると言える．ただし，これは再現率が高いためであり適合率は 0.593 とあまり高くないことに注意する必要がある．またブログに関しては F 値 0.481 と低い値になっている．しかし全体的にみるとコンテンツ構造を使った場合に比べて優れた結果になっている．そこで DBG によってどのようなクラスタが出力されているのかを次に検証する

DBG アルゴリズムによる出力の結果の例を表 3 に示す．ブログページ数は出力されたクラスタ内に含まれるラベル無しブログページの数であり，スプログ比，ブログ比は出力されたページ内の中のブログページとスプログページの割合を示している．表 3 から DBG によって抽出されるクラスタに関しては大部分がスプログか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になっていることが分かる．特に一部のクラスタに関してはスプログ割合が 96.49%，ブログ割合が 85%と非常に高い割合のコミュニティが発見されている．このような精度の高いクラスタがあるため全体の分類精度が高くなっていると考えられる．ただし，他のクラスタに関してはそこまで高い精度ではなく一部のコミュニティの精度が高い状態である．これらの結果から一部の良い精度のクラスタを選択した場合にはそれなりに高い精度となるがその他の場合には低い精度の分類となっていることが言える．このため，分類精度が一定せず全体的に見るとばらつきが多く安定した精度を得られている訳ではないと考えられる．

表 3　DBG で得られるクラスタの例

ブログページ数	スプログ比	ブログ比
392 ページ	59.44%	40.56%
174 ページ	37.36%	62.64%
135 ページ	29.63%	70.37%
73 ページ	36.99%	63.01%
60 ページ	15%	85%
57 ページ	96.49%	3.51%
53 ページ	60.38%	39.62%
39 ページ	43.59%	56.41%
9 ページ	55.56%	44.44%
9 ページ	22.22%	77.78%

表 4　リンクとコンテンツを使った場合の結果

	リンク	コンテンツ＆リンク
平均 (正解率)	64.06%	71.02%
分散 (正解率)	20.68	3.18
標準偏差 (正解率)	4.55	1.78

最後にリンクとコンテンツ構造の相補的な利用が有用であったことを示すため，リンク構造だけを使った場合とコンテンツ構造を加えて相補的に用いた場合の分類結果の性能を表 4 に示す．表では上から正解率の平均，分散，標準偏差から二つの手法の性能を比較している．

そこで正解率の分散を見てみるとリンク構造だけを訓練データに用いた場合の分散は 20.68 である，これに対してコンテンツ情報によるクラスタリングを相補的に用いることで分散は 3.18 と大幅に向上していることが見て取れる．またコンテンツ情報を相補的に用いることによる分類結果のばらつきの向上は標準偏差の値からも明らかである．さらに正解率は平均で 64.06%から 71.02%へと 7 ポイント程度の向上が見られ，結果のばらつきと分類精度双方の観点からコンテンツ情報を相補的に扱うことで有用な結果を得ていることが言える．さらに結果に対して有意水準 5%の t-検定を行い，統計的にも有意であることを示している．

これらの結果から，リンク構造を用いたクラスタは訓練データのために利用でき．さらにコンテンツ構造を相補的に用いることで分類精度，結果のばらつきの双方を向上させることが出来ているといえる．

データ全体に関しては低い分類精度だったコンテンツクラスタがリンク構造に関して有意に働いた理由としてはリンク構造のクラスタには限られたデータしかなく，さらにある程度データが偏っているためクラスタリングをし易い状態になっていたのではないかと考えられる．このことからコンテンツ構造に関するクラスタはそのままデータセット全体に用いた場合にはあまりうまく働かないが，リンク構造に基づくクラスタ内での限られたクラスタリングに関しては有用に作用していると言える．

5. むすび

本論文ではスプログの 2 つの特徴，リンク構造とコンテンツ構造に着目し，それぞれのクラスタを見つけ相補的に用いることで学習データ収集の際の人手のコストを大幅に減らす手法を提案した．

また，実際に評価データを用い人手で評価したデータでの分類と本手法により作成した学習データで分類したときの結果を比較した．その結果ごく少数の人手で評価したデータを学習データとして用いた場合，同じ学習データを使った既存手法の分類結果と比べて著しく分類精度の向上が見られた．また提案手法は得られたクラスタが必ずしも 100%の精度をもったデータではないにもかかわらず，全てラベル付けされた学習データを用いた場合の従来手法についても正解率約 8 ポイント差にまで迫った．これらの結果から学習データを一から用意した場合と比べて，提案手法は非常に小さな評価コストでそれに近い分類精度を得ることができたといえ，クラスタリングを用いた学習データの作成について一定の有効性を示した．

また，提案手法の動作を検証するためリンク構造とコンテンツ構造に対するクラスタリングの性能を比較した．その結果リンク構造は大部分のクラスタで 70%程度のクラスの偏りが得られ，さらにコンテンツ構造を相補的に用いることで最終的な分類精度を上げ，結果の偏りを減少させることが分かった．この点からリンク構造とコンテンツ構造を相補的に用いる本手法の考えが正しかったといえる．

文　献

[1] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener, “Graph structure in the web,” In Proc. of 9th Int. WWW Conf., 2000.

[2] Drucker H., Wu D, and Vapnik V. “Supprt vector machines for spam categorization.“ IEEE-NN, pages 1048-1054, 1999.

[3] Elias Dahlhaus, David S.Johnson, Christos H.Papadimitriou, Paul D.Seymour, Mihalis Yannakakis, “The Complexity of Multiterminal Cuts,” *SIAM Journal on Computing 23*, pp. 864-894 (1998)

[4] George Karypis and Vipin Kumar(1996). METIS library. Available on WWW at URL http://glaros.dtc.umn.edu/gkhome/views/metis

[5] Pranam Kolari, Akshay Java, and Tim Finin. “Characterizing the splogosphere.” *Proceedings of the 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 15th, World Wide Web Conference, May 2006.

[6] Pranam Kolari, Akshay Java, Tim Finin, Tim Oates, and Anupam Joshi. “Detecting spam blogs: A machine learning approach.” Ph.D. Dissertation, Dec 2007.

[7] Pranam Kolari, Tim Finin, Akshay Java, and Anupam Joshi. “SVMs for the Blogosphere: Blog Identification and Splog Detection.“ *In Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs,* pages 92-99, March 2006.

[8] P. K. Reddy and M. Kitsuregawa, “An approach to relate the Web communities through bipartite graphs,” In Proc. of 2nd Int. Conf. on Web Information Systems Engineering, 2001.

[9] Salton, G and M, J, McGill: “Introduction to Modern Information Retrieval,” McGraw-Hill(1983)

[10] S. R. Kumar, P. Raphavan, S. Rajagopalan and A. Tomkins, “Trawling the Web for emerging cyber communities,” In Proc. of 8th Int. WWW Conf., 1999.

[11] Yu-Ru Lin, Hari Sundaram, Yun Chi, Junichi Tatemura and Belle L. Tseng.“Splog detection using self-similarity analysis on blog temporal dynamics,” In AIRWeb ’07: *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pages 1-8, New York, NY, USA, 2007. ACM.