DBSJ Journal, Vol.8, No.1
June 2009

ソーシャルブックマークデータを用いたWebページのクラスタリング

Web Clustering Using Social Bookmarking Data

柳本 豪一♥　　吉岡 理文♦
大松 繁♦

Hidekazu YANAGIMOTO
Michifumi YOSHIOKA　Sigeru OMATU

本論文ではソーシャルブックマークデータに対して PLSI を適用し、Webページをクラスタリングする手法を提案する。ソーシャルブックマークサービスは、ユーザが興味のある Web ページを登録して他者に公開するサービスである。また、登録した Web ページにはタグと呼ばれる任意のキーワードを付加することができる。ソーシャルブックマークを利用した従来研究は Web ページのランキングに応用するものが多かった。本論文ではソーシャルブックマークデータを用いてクラスタリングを行うことで、多数のユーザの興味に基づいた Web クラスタリングを実現する。具体的にはクラスタリング手法として PLSI を用いて、Web ページのクラスタリングを実現する。評価実験として Buzzurl データを用い、主成分分析によるクラスタリング結果と比較することで提案手法の有効性を確認した。特に、提案手法は単一のトピックとしてまとまったクラスタが実現できることが確認できた。

In this paper we propose a clustering method from social bookmark data using PLSI. A social bookmark service is one of many social services and has features that many social bookmark users register Web pages and everybody can browse them. Almost all previous works use social bookmark data to reevaluate ranking from search engines but did not use them to group Web pages. We propose a method grouping Web pages with PLSI using a social bookmark service, Buzzurl. We confirmed that the proposed method classified Web pages appropriately and was superior to a previous work, PCA. Especially clusters with the proposed method have less variance than clusters with PCA.

1. はじめに

近年、新しい情報共有サービスとしてソーシャルサービスが注目を浴びている。ソーシャルサービスとして、海外のFlickr[1]やdelicious[2]、国内のはてな[3]やBuzzurl[4]などが有名である。これらのサービスは利用者が自由にタグ（キーワード）をつけて情報を登録し、その情報を公開する点に特徴がある。このため、利用者間で情報が共有され、新しい情報や有益な情報を得るための情報源として注目を浴びている。しかし、ユーザがタグに任意の単語を用いることができるため、タグに表記のゆれやユーザごとにタグの意味的な違いが発生し、効率的に必要な情報を発見できない一因となっている。このため、タグに依存しない分類手法を開発する必要がある。

本論文ではソーシャルサービスの1つであるソーシャルブックマークに対して、PLSI (Probabilistic Latent Semantic Indexing)[1]を用いたWebページのクラスタリング手法を提案する。ソーシャルブックマークサービスは個々のユーザが管理するブックマークを公開するソーシャルサービスの一つである。このため、個々のユーザの興味に基づいてフィルタリングされた情報が登録されていると考えられる。本手法では、タグを用いずソーシャルブックマークデータをユーザからWebページへのリンクとみなし、データ全体を2部グラフとして表現する。そして、この2部グラフを解析することにより、Webページのクラスタリングを実現する。ソーシャルブックマークデータより作成された2部グラフを解析する手法として本論文ではPLSIを用いる。PLSIは潜在的な属性によりデータが生成されていると仮定しており、本論文では潜在的な属性によりWebページをクラスタリングする。従来はソーシャルブックマークサービスへのデータ登録日時や登録しているユーザ数を用いたランキング手法が主な研究であるが、本論文ではクラスタリング手法に応用しており、異なったアプローチを取るものである。

以下では、従来研究についてまとめ、本論文の位置づけを明確する。次に、提案手法であるクラスタリング手法について述べ、評価実験を行う。最後に今後の課題について述べる。

2. 従来研究

従来研究としてソーシャルブックマークを用いた研究とネットワーク解析手法に関する研究を紹介する。ソーシャルブックマークを用いた研究では、ソーシャルブックマークデータの特徴、Webページのランキング応用研究について紹介する。また、ネットワーク解析に関する研究では一般的なネットワーク解析手法やネットワークを対象としたクラスタリング手法に加え、PLSIのネットワーク解析への応用研究を紹介する。以上の従来研究の紹介を通して、本論文の位置づけを明確にし、本手法の独自性を明らかにする。

ソーシャルブックマークに関する研究はFolksonomy[2]、複雑ネットワーク[3]、集合知[4]の研究を基礎としている。具体的にソーシャルブックマークサービス自体を解析した研究としてはGolderらの研究[5]、Hothoら研究[6]がある。これらの研究ではdel.icio.usからデータを収集し、ソーシャルブックマークデータの解析を行っている。解析結果としてユーザのWebページの登録数など様々なデータにスケールフリー性[7]が見られると報告している。また、Webページの登録数の時間変化に一定の傾向があると報告している。

イントラネット内でソーシャルブックマークサービスを行い解析する研究[8][9]もある。これらではイントラネット内でのソーシャルブックマークサービスが新しい情報を共有し伝搬するために有効であると報告されている。また、タ

♥ 正会員　大阪府立大学大学院工学研究科
hidekazu@cs.osakafu-u.ac.jp
♦ 非会員　大阪府立大学大学院工学研究科
{yoshioka, Omatu}cs.osakafu-u.ac.jp

[1] http://www.flickr.com
[2] http://delicious.com
[3] http://www.hatena.ne.jp
[4] http://buzzurl.jp

グが共有されることで使われるタグの総数が収束することが報告されている。以上の従来研究から、(1) ソーシャルブックマークデータがスケールフリー性を持つこと、(2) ソーシャルブックマークサービスを利用するユーザは相互に影響を及ぼすことが明らかとなっている。

以上はソーシャルブックマークサービス自体の特性についての従来研究であるが、ソーシャルブックマークデータを用いたサービスに関する研究[10]-[12]について述べる。文献[10]では、ソーシャルブックマークデータをユーザ、登録Webページ、タグの3つの属性を持つデータとみなし、それぞれが隠れ属性から生成されていると仮定してランキング手法を提案している。ユーザ、登録Webページ、タグなどの各属性は隠れ属性から生成されると仮定しており、この考え方はPLSIに類似したものである。ただし、通常のPLSIが2つの属性を扱っているのに比べ属性数が3つとなっており、PLSIで決定しなくてはいけないパラメータ数が従来に比べ多くなる。これはPLSIが過学習に陥り易い特性を持っている点[13]を考えると、好ましくないと思われる。また、登録Webページは必ずタグを持っている訳ではなく、タグに関する情報が得られない場合も多く発生する点も問題である。文献[11]ではWebページの登録情報からネットワークを構成し、HITSアルゴリズムによりランキングを行っている。これはソーシャルブックマークデータをネットワークとして扱っている一例である。文献[12]では登録Webページがどれだけのユーザにより登録されているかに着目してランキング結果の修正を行っている。具体的には、リンク構造より決定されたWebページのランキングを修正することで、ユーザの注目度を考慮したランキングを実現している。これらの手法はWebページをランキングすることが目的であり、そのままクラスタリングに応用することはできず、本論文のWebページのクラスタリングとは異なったアプローチを取る研究である。

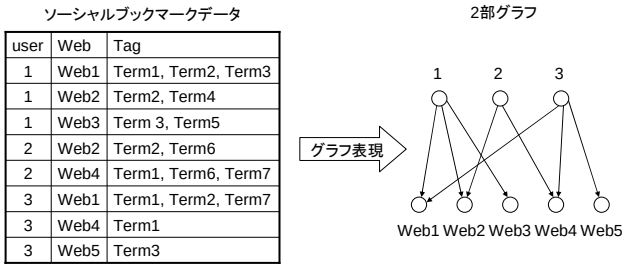
ネットワーク解析に関する従来研究としてPageRank[14]とHITS[15]があげられる。これらの手法はリンク構造を用いてWebページを評価している。具体的には接続行列の最大固有値の固有ベクトルを用いてWebページの重要度を決定している。一方、行列分解手法は確率的な枠組みのもとで再構築されており、代表的なものとしてPLSIがある。PLSIは情報検索でよく用いられるLSI(Latent Semantic Indexing)[16]を確率的枠組みでとらえ直したものであり、隠れ属性を用いて行列を分解する。CohnらはHITSとPLSIを組み合わせたPHITS[17]を提案し、複数のトピックを含むデータからクラスタリングを行いつつランキングが可能であることを報告している。これより、PLSIはネットワーク解析に有効であると思われる。今、ソーシャルブックマークデータを2部グラフと見なしてPLSIを用いて解析することで、Webページをクラスタリングすることが可能である。

一方、ネットワークの接続構造を用いた手法として、GN法[18]、Newman法[19]、CNM法[20]などがある。これらの手法ではbetweennessやmodularityを用いてネットワーク構造からクラスタリングを実現している。

3. 提案手法

本章ではソーシャルブックマークデータを用いた Web ページのクラスタリングの詳細について説明する。

ソーシャルブックマークデータとはユーザ、そのユーザが登録した Web ページ、タグなどからなるデータである。したがって、上記の 3 つの属性を一つの組としたものがそのデータである。さらに、ソーシャルブックマークでは同一の Web ページが複数回登録されることはないため、上記の組と



user	Web	Tag
1	Web1	Term1, Term2, Term3
1	Web2	Term2, Term4
1	Web3	Term 3, Term5
2	Web2	Term2, Term6
2	Web4	Term1, Term6, Term7
3	Web1	Term1, Term2, Term7
3	Web4	Term1
3	Web5	Term3

図 1　ソーシャルブックマークデータのグラフ表現
Fig.1　Graph representation of social bookmarking data

なったデータは属性値が必ず異なったものとなっている。今、ユーザがある Web ページを登録した行為をユーザから Web ページへのリンクと見なし、ソーシャルブックマークデータを 2 部グラフと見なす。具体的な 2 部グラフの構成方法を図 1 に示す。上記のように作成することで、ユーザの集合と Web ページの集合からなる 2 部グラフを構成することができる。次に、ソーシャルブックマークから構成された 2 部グラフを以下の手順により行列 N として表現する。

$$\mathbf{N} = \begin{cases} n_{ij} = 1 & (\text{ユーザ} u_i \text{が} Web\text{ページ} r_j \text{を登録}) \\ n_{ij} = 0 & (\text{その他}) \end{cases}$$

ただし、n_{ij}は行列 N の(i,j)成分を表す。

上記の説明では2部グラフを構成するためにタグを用いていない。2 部グラフの構成にタグを用いない理由は、(1)タグの表記の揺れやユーザ間での意味的な相違の影響により、ユーザと Web ページ間の不要なリンク生成を避ける、(2) タグが設定されていない Web ページへの対応問題を回避するためである。タグの利用状況については、実験結果で詳細を述べる。一方、独立にタグを扱うと、ソーシャルブックマークデータの属性数を増やすこととなり、PLSI で決定すべきパラメータ数が増加する。このような問題の原因は、Web ページに対してタグを複数設定することができ、タグに利用できる語彙が制限されていないためである。以上の点を考慮して、本論文ではタグを属性に用いない。他にも登録日を属性とすることも考えられるが、情報の新規性を考慮したクラスタリングが目的ではないため、属性から除外した。

ソーシャルブックマークより作成した行列 N を用いて PLSI でクラスタリングを行う。PLSI ではユーザ u_iと Web ページ r_jの同時確率を $P(u_i, r_j)$とし、これを隠れ属性 z_kを用いて分解する。

$$P(u_i, r_j) = P(z_k) P(u_i \mid z_k) P(r_j \mid z_k)$$

ここでは aspect model[18]と同様の生成モデルを用いており、ユーザ u_iと Web ページ r_jは隠れ属性 z_kの条件のもとで独立（条件付き独立）であると仮定している。Web ページをクラスタリングするため、ある隠れ属性において高い確率を持つ Web ページの集合を一つのクラスタと見なし、Web ページのクラスタリングを実現する。つまり、$P(r_j \mid z_k)$に注目し、大きな値を持つ Web ページを同じクラスタとすることでクラスタリングを行う。

観測値は多項分布に従っていると考え、Web クラスタリングで用いる条件付き確率 $P(r_j|z_k)$を求める。具体的には以下の対数尤度 L を用いて最尤推定値を求めることで実現する。

$$L = \sum_{i,j} n_{ij} \log \sum_{k} P(z_k) P(u_i \mid z_k) P(r_j \mid z_k)$$

上記の式は Jensen 不等式を用いることで、対数内の和を対数の外に移すことができる。具体的には、以下のように書き換えることができる。

$$L \geq \sum_{i,j,k} n_{ij} P(z_k \mid u_i, r_j) \log \frac{P(z_k) P(u_i \mid z_k) P(r_j \mid z_k)}{P(z_k \mid u_i, r_j)}$$

上式では新しく $P(z_k|u_i,r_j)$を導入している。EM アルゴリズムでは E-step において $P(z_k|u_i,r_j)$を決定し、個々の条件付き確率は M-step で推定する。具体的な更新式は以下となる。

・E-step

$$P(z_k \mid u_i, r_j) = \frac{P(z_k) P(u_i \mid z_k) P(r_j \mid z_k)}{\sum_k P(z_k) P(u_i \mid z_k) P(r_j \mid z_k)}$$

・M-step

$$P(u_i \mid z_k) = \frac{\sum_j n_{ij} P(z_k \mid u_i, r_j)}{\sum_{i,j} P(z_k \mid u_i, r_j)}$$

$$P(r_j \mid z_k) = \frac{\sum_i n_{ij} P(z_k \mid u_i, r_j)}{\sum_{i,j} P(z_k \mid u_i, r_j)}$$

$$P(z_k) = \frac{\sum_{i,j} n_{ij} P(z_k \mid u_i, r_j)}{\sum_{i,j} n_{ij}}$$

上記の EM アルゴリズムを実行するためには、各確率の初期値が必要である。初期値の設定として $P(u_i|z_k)$は乱数、$P(z_k)$と $P(r_j|z_k)$は一様分布を用いる。

4. 実験

株式会社 EC ナビが提供するソーシャルブックマークサービスである Buzzurl の 2005 年 10 月から 2008 年 10 月までの登録データを用いる。まず、ソーシャルブックマークデータの特徴について検討を行い、提案手法を用いたクラスタリング結果について報告する。クラスタリング結果の検討として、同一のトピックでクラスタリングしているかという観点と、スパム Web ページの影響について検討する。ここで、スパム Web ページとはソーシャルブックマークサービスの提供者が好ましくないと判断する Web ページのことであり、年齢制限が必要な Web ページなどがその一例である。

4.1 ソーシャルブックマークデータの特徴

表 1 に Buzzurl データの全体構成を示す。このデータは 2005 年 10 月から 2008 年 10 月までの全データが含まれている。最後の Data 項目は図 1 に示したようなユーザ、Web ページ、タグを組としたデータ数である。このデータを用いて、Web ページの登録数の分布（図 2）、ユーザの登録数の分布（図 3）を表す。このグラフより、両分布はスケールフリー性を表しており、多くの Web ページはごく少数のユーザからしか登録されておらず、ごく少数の Web ページしか登録していないユーザが多くを占めていることが分かる。

次に Web ページに設定されたタグ数について調べる。図 4 に示すように、全データの約 1/4 においてタグが付けられていない。また、図 5 より、タグが設定されていない Web ページが毎月一定数登録されていることが分かる。以上より、Web ページにタグが付加されていないものが多く含まれており、タグを利用できない場合が多く発生する。

表 1　Buzzurl データの構成
Fig.1　Contents of Buzzurl data.

Unique User	25,597
Unique Web Page	864,574
Unique Tag	352,015
Data	1,626,869

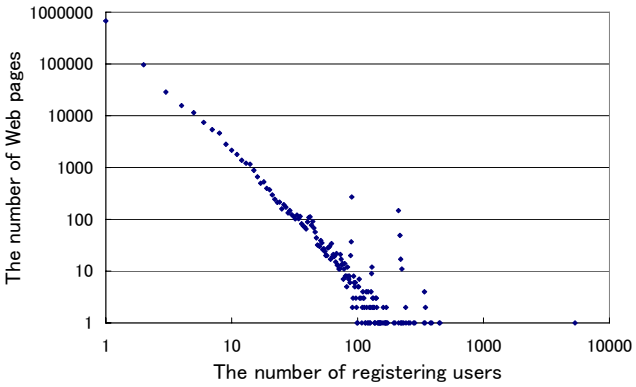


図 2　Web ページの登録数の分布
Fig.2　Web Page Distribution with respect to the number of registering users.

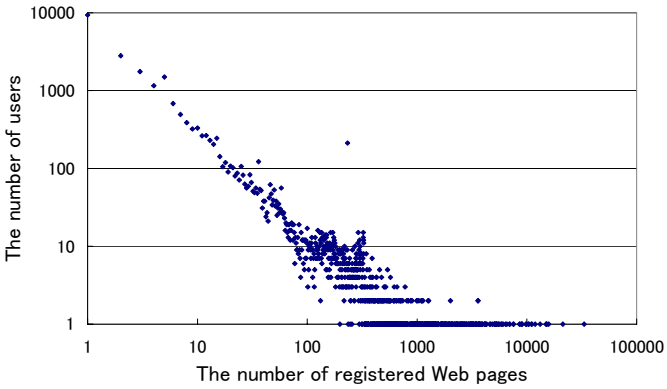


図 3　ユーザの登録数の分布
Fig.3　User Distribution with respect to the number of registered Web pages.

4.2 Web ページのクラスタリング

PLSI による Web ページのクラスタリングを行い、クラスタリング結果の検討を行う。特に、同一のトピックでクラスタがまとまっているか検討を行う。また、入力データに含まれるスパム URL の影響についても検討する。

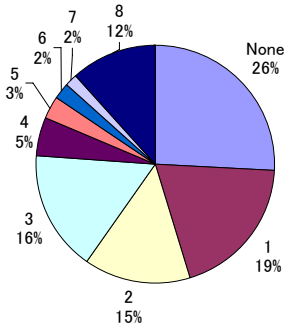


図 4　タグ数の分布
Fig.4　Tag Distribution with respect to Web pages.

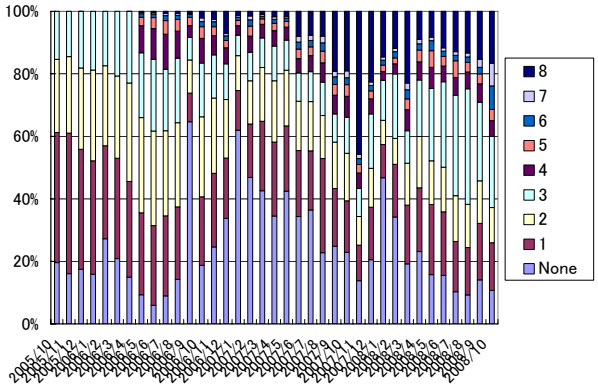


図 5　月別の登録タグ数の変化
Fig.5　The number of monthly registered tag.

4.2.1 実験環境

Web ページのクラスタリングを行うため、利用するソーシャルブックマークデータの数を絞り込んで入力データを作成する。これは(1)多数のユーザの評価に基づいたクラスタリングを行い、(2)計算量を減らすためである。本実験では登録 URL は最低 20 人のユーザから登録されているものとした。また、10 件以上の URL を登録しているユーザのみを対象とした。これらの閾値は計算量に注目して数値を決定した。これより、2,256 人のユーザ、5,248 件の Web ページを用いた。したがって、2,256 ×5,248 の行列が入力行列 N となる。比較手法として主成分分析 (PCA) を用いる。なぜなら、PCA と特異値分解、特異値分解と PLSI の類似性のためである。PCA では行列 N^TN の固有値分解により得られた固有ベクトルでクラスタリングする。ここで・^{T}は行列の転置を表す。

次に、隠れ属性の数について検討する。隠れ属性数は PLSI が構成する確率モデルの複雑さに影響を与えるため、適切な数を設定する必要がある。あらかじめデータがいくつのクラスタを構成するか分かっている場合を除いて、適切な隠れ属性数をあらかじめ決定することは困難である。これは PCA において適切な主成分数を決定する困難さと同様である。PCA では寄与率を用いて主成分数を決定することが多い。よって、本論文では寄与率が 0.6 以上となるように主成分数を決めることとし、その主成分数を隠れ属性数とする。予備実験より、主成分数が 30 で寄与率が 0.6 を超えたため、以下では主成分数、隠れ属性数を 30 とした。

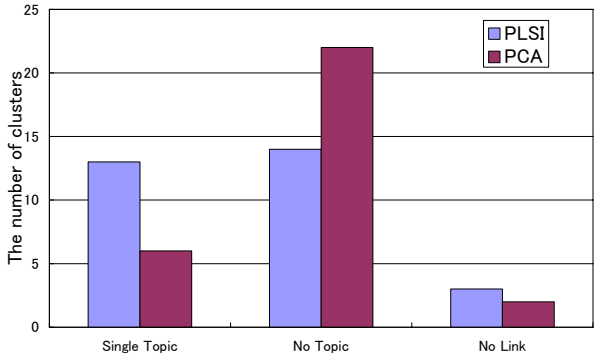


図 6　クラスタリング結果
Fig.6　Clustering with PLSI and PCA

4.2.2 クラスタリング結果

クラスタリング結果の評価としては、それぞれのクラスタに含まれる上位 5 件の Web ページが同一のトピックおよび同一サービスであるかどうかを手作業で判断した。判断基準として Web ページのタイトルなどから同一のトピックを扱っているか、同一サービスであるかを判断した。5 件の Web ページ中 4 件以上で同じトピックおよびサービスを扱っている場合に単一のトピックを扱っていると判断した。結果を図 6 に示す。クラスタリング結果のうち、クラスタに含まれるすべての Web ページへのリンク先が消滅していて内容を確認できないものがあったため、それは個別に表している。

この結果から分かるように、提案手法の方がトピックにより Web ページを分類できていた。トピックが判断できないものにはニュースサイトやブログサイトなどが混在している場合が多く、多岐な話題を含んでいるため、トピックとしてまとまっていると判断することが困難であった。

具体的にクラスタ結果について検討する。表 2 にクラスタリングの結果の一部を示す。Latent Attribute 9 と Principal Component 22 は PLSI および PCA において単一トピックでまとまったと判断した例である。Latent Attribute 9 は通信販売での食品のランキングであり、Principal Component 22 はデニム生地の衣類のランキングである。ともに商品のランキングから構成されるページであり、クラスタ内で関連した商品情報を扱っている。提案手法のみから得られたクラスタの例を Latent Attribute 15 で説明する。このクラスタはポータルサイトをまとめたものであると判断できる。関連したサービスをソーシャルブックマークに登録するユーザが多いため、このようなクラスタを見つけることができたと考えられる。上位 5 件以下の Web ページを表 4 に示すが、同様にポータルサイトやトップページが多く含まれている。Latent Component 15 と Principal Component 30 は似た Web ページを含むクラスタの例である。このクラスタに含まれる Web ページはギャンブルに関するものである。一般的に、これらはソーシャルブックマークサービスを提供する側からは好ましくないものであり、特定のユーザが大量に登録したものであると考えられる。これらの Web ページはスパムメールに似ており、スパム web ページの一例である。

4.2.3 スパム Web ページのクラスタリングへの影響

多くのユーザによって登録された Web ページには有益な情報を含む Web ページ以外にスパム Web ページも含まれ

表 3　クラスタリング例
Fig.3　Examples of clustering results.

Latent Attribute 9 in PLSI
http://www.lifeangel.co.jp/ http://datyounoniku.sblo.jp/ http://odenkan.sblo.jp/ http://satsuporonomen.sblo.jp/ http://yanbarusimabuta.sblo.jp/
Principal Component 22 in PCA
http://deninoiihuku.sblo.jp/ http://denibes.sblo.jp/ http://iimonosaro.sblo.jp/ http://denisaro.sblo.jp/ http://guranohakimono.sblo.jp/
Latent Attribute 15 in PLSI
http://www.yahoo.co.jp/ http://www.mdn.co.jp/ http://twitter.com/ http://b.hatena.ne.jp/ http://www.google.co.jp/ig?hl=ja/
Latent Attribute 18 in PLSI
http://www.onlinesanctuary.com/bookmaker/ http://xn--lckh3dvdtc8ib.jp/ http://xn--kck6a0a2002a9zu255c8vm.jp/ http://xn--7rs178bg0js23a.jp/ http://www.vos-net.com/
Principal Component 30 in PCA
http://xn--eckaqff5e5fob9rsdc5271n85ub.jp/ http://xn--y9qv79bor2athh.jp/ http://www.onlinesanctuary.com/bookmaker/ http://xn--7rs178bg0js23a.jp/ http://xn--kck6a0a2002a9zu255c8vm.jp/

表 4　Latent Attribute 15 の後続 Web ページ
Fig.4　Following Web pages in Latent Attribute 15.

http://www.flickr.com/ http://buzzurl.jp/ http://www.gyao.jp/ http://www.youtube.com/ http://www.apple.com/

ている。一般的にスパム Web ページは多くのユーザから登録されている形を取っているため、登録数を元に入力データの絞り込みを行ったため避けることは難しい。ここではクラスタリング結果とスパム Web ページの関係について検討する。

まず、入力データに含まれるスパム Web ページの割合を表 5 に示す。この情報は Buzzurl データに含まれている情報であり、スパム Web ページかどうかの評価は人が Web ページを閲覧して判断を行っている。具体的にはランキングだけの情報が少ない Web ページやギャンブルなどの Web ページがスパム Web ページと分類されている。SPAM-like URL はまだ人手では判断が行われていないが、過去の登録 Web ページの履歴からスパム Web ページである疑いが高い Web ページである。以下では SPAMURL と SPAM-like URL をスパム Web ページであるとみなして以下では検討を行う。

入力データの分布から分かるように、入力データの約半分がスパム Web ページである。スパム Web ページが登録数を

表 5　入力データにおけるスパム Web ページの分布
Fig.5　SPAM Web pages in input data.

Not SPAM URL	2,711
SPAM URL	1,715
SPAM-like URL	822

表 6　スパム Web ページで構成されたクラスタ
Fig.6　Clusters including SPAM Web pages.

PLSI	16
PCA	22

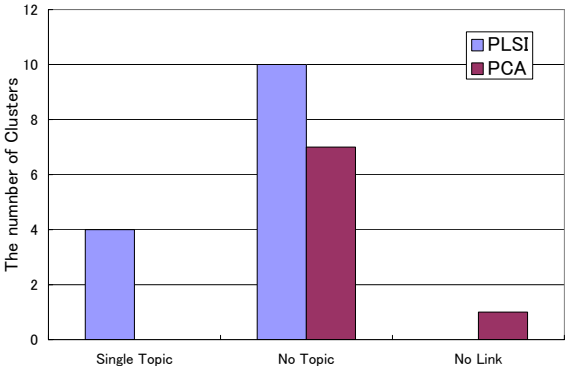


図 7　スパム Web ページを除いたクラスタリング結果
Fig.7　Clustering with PLSI and PCA without SPAM Web pages.

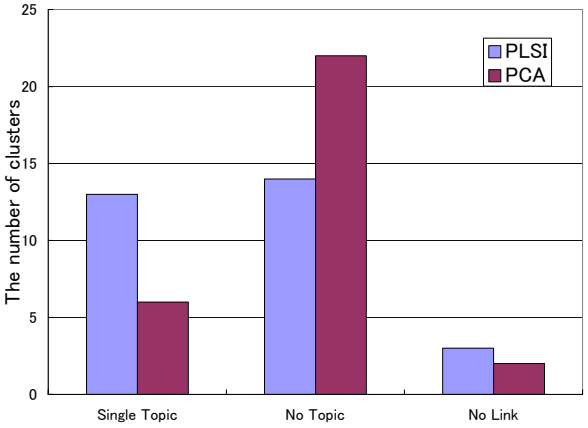


図 8　スパム Web ページのみのクラスタリング結果
Fig.8　Clustering with PLSI and PCA for SPAM Web pages.

多くすることを目指していることから仕方がない。クラスタリング結果を推薦などに応用することを考えると、クラスタリング結果においてスパム Web ページでない Web ページから構成されるクラスタが多く得られることが好ましい。今、クラスタリング結果のうち、上位五件の Web ページがすべてスパム Web ページであるクラスタをスパムクラスタとみなして、スパムクラスタ数を表 6 に示す。この結果より、提案手法の方がスパム Web ページ以外の Web ページをクラスタとしてまとめることができている。PCA で悪くなった理由としては、スパム Web ページは特定のユーザが登録していると考えられるので、登録しているユーザと登録していないユーザが明確に分かれ、分散が大きくなるため精度が悪くなったと考えられる。また、上位 5 件にスパム Web ページ

を含むクラスタは、続く上位 6 位から 10 位の Web ページはすべてスパム Web ページであった。以上より、両手法ともスパム Web ページをまとめることはできていると思われる。図 7 にスパム URL を除いたときの分類結果を示す。これより、提案手法の方が入力データにスパム URL が混在していた場合も有益な情報を持つ Web ページをクラスタリングできることが確認できた。また、図 8 より提案手法の方がスパム Web ページを内容でまとめることができている。

5. おわりに

ソーシャルブックマークデータを用いた Web ページのクラスタリングを提案した。実験より、タグを利用せず、単一トピックのクラスタが構成可能であることを確認した。

一方、隠れ属性数の決定についてはオープンな問題として残っている。情報量基準などを用いることにより、適切な隠れ属性数が決定できるか検討する必要がある。そして、タグを用いない本手法では分類性能が約 50%である。このため、タグを組み合わせることで改善を行う必要がある。この際、Latent Dirichlet Allocation などの PLSI 以外の手法を利用することも考える必要がある。また、他のクラスタリング手法との比較することにより本手法の検討を行う必要がある。

[謝辞]

株式会社 EC ナビから提供された Buzzurl ソーシャルブックマークデータを用いて本実験は行われている。ここに記して謝意を表します。

[文献]

[1] T. Hofmann: “Probabilistic Latent Semantic Indexing", Proc. of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval, (1999).

[2] A. Mathes: “Folksonomies - Cooperative Classification and Communication Through Shared Metadata", http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

[3] D. J. Watts and S. H. Strogatz: “Collective Dynamics of `small-world' Networks", Nature, Vol.393, pp.440--442, (1998).

[4] 大向一輝: “Web2.0 と集合知", 情報処理, Vol.47, No.11, pp.1214--1221, (2006).

[5] S. Golder and B. A. Huberman: “The Structure of Collaborative Tagging Systems", Journal of Information Science, Vol.32, No.2, pp.198--208, (2006).

[6] A. Hotho, R. Jaschke, C. Schmitz and G. Stumme: “Information Retrieval in Folksonomies: Search and Ranking", Proc. of Third European Semantic Web Conference, pp.411--426, (2006).

[7] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley and Y. Aberg: “The web of human sexual contacts", Nature, Vol.411, pp.907--908, (2001).

[8] L. Damianos, J. Griffith and D. Cuomo: “Onomi: Social Bookmarking on a Corporate Intranet", International Collaborative Web Tagging Workshop at WWW2006, (2006).

[9] D. R. Millen, J. Feinberg and B. Kerr: “Dogear: Social Bookmarking in the Enterprise", Proc. of the SIGCHI Conference on Human Factors in Computing Systems, pp.111--120, (2006).

[10] X. Wu, L. Zhang and Y. Yu: “Exploring Social Annotations for the Semantic Web", Proc. of the 15th International Conference on World Wide Web, pp.417--426, (2006).

[11] H. Wu, M. Zubair and K. Maly: “Harvesting Social Knowledge from Folksonomies", Proc. of the 17th Conference on Hypertext and Hypermedia, pp.111--114, (2006).

[12] 山家雄介, 中村聡史, A. Jatowt, 田中克己: “Web 検索のランキング精度向上のためのソーシャルブックマークの利用", DEWS2007, (2007).

[13] D. M. Blei, A. Y. Ng and M. I. Jordan: "Latent Dirichlet Allocation", Journal of Machine Learning Reserach, Vol.3, pp.993--1022, (2003).

[14] S. Brin and L. Page: “The Anatomy of a Large-scale Hypertextual Web Search Engine", Computer Networks and ISDN Systems, Vol. 30, No.1-7, pp.107--117, (1998).

[15] J. M. Kleinberg: “Authoritative Sources in a Hyperlinked Environment", Journal of the ACM, Vol. 46, No.5, pp.604--632, (1999).

[16] S. Deerwester and S. Dumais, G. W. Fumas, T. K. Landauerand R. Harshman: “Indexing by Latent Semantic Analysis", Journal of the ACM for Information Science, Vol. 41, No. 6, pp.391--407, (1990).

[17] D. Cohn and H. Chang: “Learning to Probabilistically Identify Authoritative Documents", Proc. of the 17th International Conference on Machine Learning, (2000).

[18]M.E.J.Newman and M.Girvan: “Finding and evaluating community structure in networks”, Physical Review, E 69, 026113(2004)

[19]M.E.J.Newman: “Fast algorithm for detecting community structure in networks”, Physical Review, E 69, 066133(2004)

[20] Aaron Clauset, M.E.J.Newman and Cristopher Moore: “Finding community structure in very large networks”, Physical Review, E 70, 066111 (2004)

[21] T. Hofmann, J. Puzicha and M. I. Jordan: “Unsupervised Learning from dyadic data", In Advances in Neural Information Processing Systems, Vol. 11, (1999)

柳本 豪一 Hidekazu YANAGIMOTO

1972 年生．1996 年大阪府立大学大学院工学研究科博士前期課程修了．同年日本電気株式会社入社．2000 年より大阪府立大学工学部助手となり現在に至る．情報検索，進化型計算手法に関する研究に従事．工学(博士)．情報処理学会，電子情報通信学会，人工知能学会，計測自動制御学会，システム制御情報学会会員，電気学会，日本データベース学会会員．

吉岡 理文 Michifumi YOSHIOKA

1968 年 12 月 10 日生. 1996 年 3 月東京大学大学院工学研究科博士課程修了．同年 4 月大阪府立大学工学部助手，1998 年 11 月同講師となり現在に至る. 画像処理等の研究に従事．工学博士，情報処理学会会員，計測自動制御学会会員．

大松 繁 Sigeru OMATU

1974 年大阪府立大学大学院博士課程修了．同年徳島大学工学部情報工学科助手．1988 年同知能情報工学科教授。1995 年大阪府立大学工学部情報工学科教授となり，現在に至る。1991 年電気学会論文賞，1995 年計測自動制御学会論文賞，1996 年市村賞受賞。ニューラルネットワークの研究に従事。IEEE Trans. on Neural Network の Associate Editor。システム制御学会，計測自動制御学会，IEEE 会員。工学博士