ソーシャルブックマークデータを用いたWebクラスタリング

Web clustering using social bookmarking data

柳本 豪一*1　　大松 繁*1
Hidekazu YANAGIMOTO　Sigeru Omatu

*1大阪府立大学
Osaka Prefecture University

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ブックマークサービスが注目を浴びている。ソーシャルブックマークサービスとして、delicious*1、はてな*2、Buzzurl*3 などが存在する。これらのサービスでは、ユーザが気に入った Web ページを登録し、その Web ページに対してタグを付けて情報を公開することができる。このタグは登録したユーザが情報を管理したり、その他のユーザが情報を探す際に利用される。しかし、タグは任意な単語を用いることができるため、表記の揺れが存在し、ユーザ間で意味的な相違が存在するため、他のユーザの情報を効率的に探すことが困難である。

従来研究ではソーシャルブックマークデータを用いた Web ページのランキングを行う手法が主であった。文献 [Wu X. 06] ではソーシャルブックマークデータをユーザ、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、ランキングを行っている。これは PLSI[Hofmann 99-1] に類似した手法であるが、通常 PLSI は 2 つの属性を扱っているのに比べ、属性数が多くなっている。このため、PLSI が決定しなくてはならないパラメータ数が増加し、過学習に陥りやすいと考えられる。文献 [Wu H. 06] ではソーシャルブックマークデータからネットワークを構成し、HITS アルゴリズム [Kleinberg 99] を用いてランキングを行っている。文献 [Yanbe 07] ではユーザからの登録数を用いてサーチエンジン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修正を実現している。このように従来研究ではソーシャルブックマークデータを用いることで、サーチエンジンとは異なる観点からのランキングを作成することが主であった。しかし、ソーシャルブックマークサービスを用いて関連した情報を集める場合を考えると、ランキングより登録された Web ページを整理できるクラスタリングが有効な場合がある。例えば、特定のクラスタに分類された Web ページを推薦することで関連情報を提供することが可能である。

本論文ではソーシャルブックマークデータをクラスタリングし、関連した情報をまとめることを目的とし、ユーザの情報発見を支援することを目指す。まず、ユーザが Web ページを登録する行為をユーザから Web ページへのリンクとみなし、ソーシャルブックマークデータをネットワークとして表現する。そして、得られたネットワークを PLSI(Probabilistic Latent Semantic Analysis) を用いて解析し、関連した Web ページを同一のクラスタリングに分類する。提案手法の性能を検討するため、Buzzurl ブックマークサービスのデータを用いた評価実験を行い、提案手法がノイズが少ないクラスタを構成できることを確認した。

連絡先: 柳本 豪一，大阪府立大学大学院工学研究科，堺市学園町 1-1，hidekazu@cs.osakafu-u.ac.jp

∗1 http://delicious.com
∗2 http://www.hatena.ne.jp
∗3 http://buzzurl.jp

2. 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 クラスタリング手法について述べる。まず、ソーシャルブックマークデータをネットワークで表現する方法について述べる。次に、PLSI を用いた Web クラスタリングについて説明する。

2.1 ソーシャルブックマークの行列表現

ソーシャルブックマークデータはユーザ、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。本手法ではユーザと登録 Web ページに着目して、Web ページのクラスタリングを行う。まず、ユーザが Web ページを登録するという行為をユーザから Web ページへのリンクとみなし、2 部グラフを構成する。次に、この 2 部グラフから以下の手順により行列 $\mathbf{N}$ を作成する。

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

ここで、n_{ij} は行列 $\mathbf{N}$ の (i, j) 成分を表す。

本手法ではソーシャルブックマークデータから 2 部グラフを構成する際にタグを用いていない。これは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、関連性のない Web ページに対してリンクを生成しないためである。また、実際のデータを見るとタグが付けられていないデータも多く存在するため、利用しないこととした。

2.2 Web ページのクラスタリング

前節で作成した行列 $\mathbf{N}$ を用いて Web ページのクラスタリングを行う。クラスタリング手法として、本論文では PLSI を用いるため、PLSI について説明を行う。PLSI ではユーザ u_i と Web ページ r_j の同時確率を $\mathrm{Pf}(u_i, r_j)$ とし、これが隠れ属性 z_k を用いて分解を行う。

$$\Pr(u_i, r_j) = \sum_k \Pr(z_k)\Pr(u_i|z_k)\Pr(r_j|z_k) \quad (2)$$

ここで、aspect model[Hofmann 99-2] と同様な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 のもとで条件付き独立であると仮定している。以上の分解より得ら

表 1: Buzzurl データの構成

ユーザ数	25,597
Web ページ数	864,574
タグ数	1,626,869

れた $\Pr(r_j|z_k)$ に着目し、大きな値を持つ Web ページを取り出すことでクラスタリングを作成する。

上記の条件付き確率を求めるため、対数尤度 L を用いる。

$$L = \sum_{i,j} n_i j \log(\sum_k \Pr(z_k)\Pr(u_i|z_k)\Pr(r_j|z_k)) \quad (3)$$

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。

$$L \geq \sum_{i,j,k} n_{ij}\Pr(z_k|u_i,r_j) \log \frac{\Pr(z_k)\Pr(u_i|z_k)\Pr(r_j|z_k)}{\Pr(z_k|u_i,r_j)} \quad (4)$$

上式では新しく $\Pr(z_k|u_i,r_j)$ を導入している。EM アルゴリズムでは E-step において $\Pr(z_k|u_i,r_j)$ を決定し、M-step において個々の条件付き確率を推定する。具体的な更新式を以下に示す。

- E-step

$$\Pr(z_k|u_i,r_j) = \frac{\Pr(z_k)\Pr(u_i|z_k)\Pr(r_j|z_k)}{\sum_k \Pr(z_k)\Pr(u_i|z_k)\Pr(r_j|z_k)}$$

- M-step

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

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

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

上記の EM アルゴリズムを実行するため、各確率の初期値が必要である。初期値としては、$\Pr(u_i|z_k)$ は乱数、その他は一様分布を用いることとする。

3. 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 の登録データを実験に用いた。まず、実験データの特徴の検討を行い、次に提案手法を用いた Web ページのクラスタリングを行う。

3.1 ソーシャルブックマークデータの検討

表 1 に全 Buzzurl データの構成を示す。このデータは 2005 年 10 月から 2008 年 10 月までの全データである。Web ページやタグはユニークな Web ページやタグの数を表している。実際のソーシャルブックマークデータはユーザ、Web ページの URL、タグなどが組み合わされたものであり、160 万以上のデータとなっている。

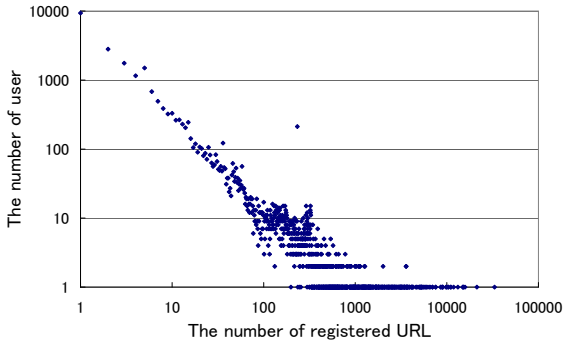

図 1: ユーザの Web ページの登録数の分布

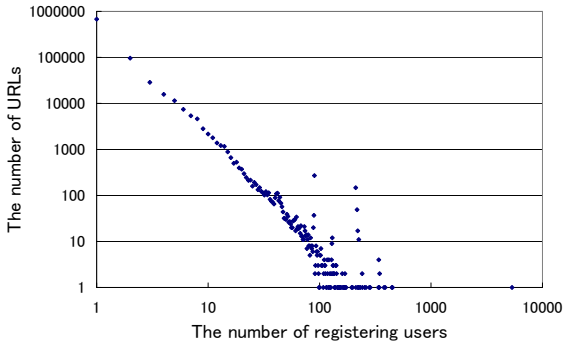

図 2: Web ページの登録数の分布

まず、ユーザの Web ページの登録数の分布について図 1 を用いて検討する。図 1 から分かるようにこの分布はスケールフリー性を有しており、大部分のユーザが 1 件の Web ページしか登録しておらず、全ユーザの約 37%を占めている。一方、Web ページの登録数の分布についても検討するため、図 2 に分布を示す。これも同様にスケールフリー性を有し、一人のユーザからしか登録されていない Web ページが約 78%を占めている。最後にタグについて検討を行う。図 3 から分かるように、スケールフリー性を有し、一度しか使われていないタグが約 60%を占めている。

以上より、ソーシャルブックマークデータはスケールフリー性をいろいろな面で有しており、特定の少数のデータのみが密に他のデータと関連していることが分かる。

次ににソーシャルブックマークデータにおけるタグの利用状況について検討する。図 4 はソーシャルブックマークデータで利用しているタグの数を表したものである。これより、約 $\frac{1}{4}$ においてタグが利用されていないことがわかる。また、図 5 より、タグが設定されていないソーシャルブックマークデータは毎月一定数登録されていることがわかる。以上の点からも、本手法ではタグを用いないこととした。

3.2 Web クラスタリング

Web クラスタリングの実験には上記のデータを全て使うのではなく、絞り込んで入力データを作成する。これは (1) 多数のユーザの判断に基づいた分類を行う、(2) 計算量を減らすためである。本手法の目的は多くのユーザの意見に基づいて Web ページを分類することであるため、少数のユーザからしか登録されていない Web ページを除くことは妥当であると考

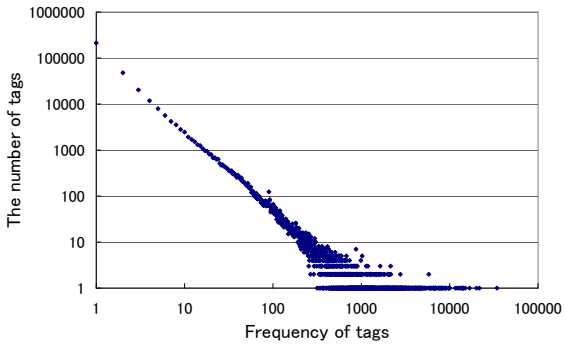

図 3: タグの出現頻度の分布

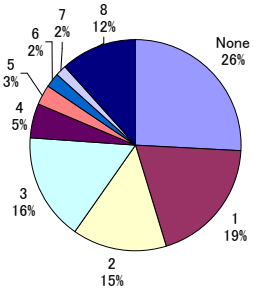

図 4: タグ数の分布

える。また、解析のアルゴリズムとして行列の分解を行うため、大規模すぎる行列は計算時間の増加に繋がるため、制限を行うこととした。本実験では最低 20 名のユーザから登録されている Web ページ、10 件以上の Web ページを登録しているユーザを入力データとして用いることとした。この条件を満たす 2,256 名のユーザ、5,248 件の Web ページを入力データとして用いた。したがって、入力データは $2,256 \times 5,248$ の行列 $\mathbf{N}$ となる。

比較手法として主成分分析を用いる。行列 $\mathbf{N}^{\mathrm{T}}\mathbf{N}$ を固有値分解することによって得られる固有ベクトルを用いて Web のクラスタリングを行う。ここで $\cdot^{\mathrm{T}}$ は行列の転置を表す。

次に隠れ属性の数について説明する。あらかじめクラスタ数が分かっている場合を除いて、適切な隠れ属性数を決定する

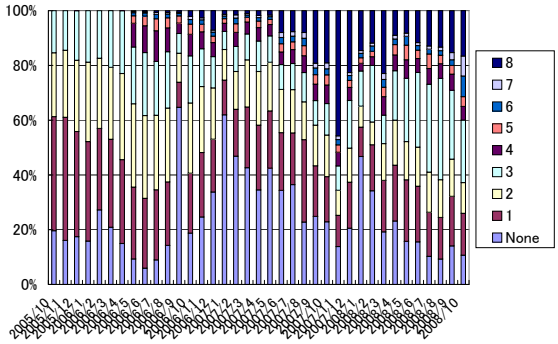

図 5: 月別の登録タグ数の変化

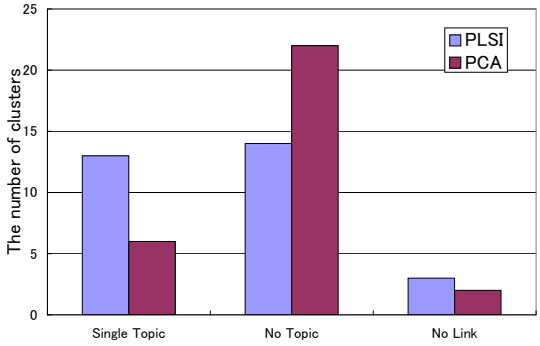

図 6: クラスタリング結果

ことは困難である。本実験では主成分分析における寄与率を用いることとし、寄与率が 0.6 を超えた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、隠れ属性数は共に 30 として実験を行った。

クラスタリング結果の評価として、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、もしくは同一のサービスであるかを手作業で判断し、正解率を求めた。これは Buzzurl データは実際にサービスに用いられているデータであり、正解のラベルはつけられていないためである。具体的には Web ページのタイトルなどから同一性を判断した。5 件中 4 件以上で同一と判断した場合に単一のトピックからクラスタが構成されていると判断した。結果を図 6 に示す。クラスタリングされた Web ページの中にはリンク先が消滅しているものがあるため、すべての Web ページのリンク先が消滅している場合は結果に含めていない。この結果より提案手法の方がトピックにより Web をクラスタリングし、ノイズが少ないことが分かる。まとまっていないと判断されたクラスタにはニュースサイトやブログサイトを含んでいるものが多く、クラスタが多岐な話題にわたっているため、まとまっているとは判断しなかった。

クラスタについて具体的に検討する。表 2 にクラスタ結果の一部を示す。Latent Attribute 9 と Pricipal Component 22 は PLSI と PCA において単一トピックから構成されていると判定された例である。前者は食べ物のランキング、後者はデニム生地の衣類のランキングである。ともにクラスタ内では関連した商品情報を扱っており、同質のデータを持つクラスタである。Latent Attribute 15 は提案手法のみで見られた結果であり、ポータルサイトをまとめたクラスタとなっている。表 3 に Latent Attribute 15 の残りの Web ページのリストを示す。これらもポータルサイトであり、さまざまなサイトの情報をまとめることができている。Latent Attribute 18 と Principal Component 30 はともにギャンブルに関する Web ページから構成されるクラスタである。これらのページはソーシャルブックマークを提供する側からトップページに掲載されることが好ましくないものである。これらをクラスタとしてまとめることにより、サービスの保守の支援を行えると考えられる。

4. まとめ

本論文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を提案した。PLSI を用いることで、ノイズが少なく関連性の高い Web ページをまとめたクラスタを構

表 2: クラスタリング結果の例

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/

成できることが評価実験より確認できた。

一方、隠れ属性数の決定に関しては問題が残っている。情報量基準などを用いることで、適切な隠れ属性数が決定できるか検討する必要がある。また、本手法ではタグを用いていないため、今後タグを組み合わせたクラスタリング手法の開発を行う予定である。

最後に、実験データを提供していただいた株式会社 EC ナビに感謝いたします。

参考文献

[Wu X. 06] Wu X., Zhang L., Yu Y.: Exploring Social Annotations for the Semantic Web, Proc. of the 15th International Conference on World Wide Web, pp.417–426(2006)

[Wu H. 06] Wu H., Zubair M., Maly K.: Harvesting Social Knowledge from Folksonomies, Proc. of the 17th Conference on Hypertext and Hypermedia, pp.111–114(2006)

[Kleinberg 99] Kleinberg J. M.: Authoritative Sources in a Hyperlinked Environment, Journal of the ACM, Vol.46, No.5, pp.604–632 (1999)

表 3: Latent Attribute 15 の後続 Web ページ

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

[Yanbe 07] 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)

[Hofmann 99-1] Hofmann T.: Probabilistic Latent Semantic Indexing, Proc. of the 22th Annual Intenational SIGIR Conference on Research and Development in Information Retrieval (1999)

[Hofmann 99-2] Hofmann T., Puzicha J., Jordan M. I.: Unsupervised learning from dyadic data, In Advances in Neural Information Processing Systems, Vol.11 (1999)