1D1-1

ソーシャルブックマークを用いたWebページの分類

Web Clustering Using Social Bookmarking Data

柳本 豪一[*1] 吉岡 理文[*1] 大松 繁[*2]
Hidekazu Yanagimoto Michifumi Yoshioka Sigeru Omatu

[*1]大阪府立大学 Osaka Prefecture University
[*2]大阪工業大学 Osaka Institute of Technology

We propose a web clustering method using social bookmarking data with dimensionality reduction regarding similarity based on their cooccurrence. In social bookmarking services a suser evaluates web pages according to his/her interests and decides whether they are registered or not. Web pages registered by a user has some relationships. Hence, all web pages make a network using their cooccurrence frequency. However, some web pages registered by the same user do not have the relationship because of users' multiple interests. We reduce such the relationship mapping web page network onto less dimensional feature space. We evaluate our proposed method using artificially generated data and Buzzurl social bookmarking data.

1. はじめに

近年、ソーシャルブックマークサービスはインターネット上の情報源として注目されている。国外では delicious[*1]、国内でははてな[*2]、Buzzurl[*3] などがソーシャルブックマークサービスの一例である。ソーシャルブックマークサービスでは、利用者が個人の興味に基づいて Web ページを登録し、その登録情報がインターネット上で公開される。この特徴のため、有益な情報 (Web ページ等) は多くの利用者の共有され、利用者集合の中で情報の選別が行われる。特に、利用者が Web ページを評価して登録を行い、利用者間で有益な Web ページが共有されるため、サーチエンジンなどの検索結果とは異なる質の高い情報を見つけることができる点が特徴である。そこで、ソーシャルブックマークデータを解析することで、利用者の評価に基づいた Web ページのクラスタリングを行うことを考える。

本論文ではソーシャルブックマークデータを用いて、Web ページを重み付きグラフで表現する。利用者が同時に登録する Web ページは何らかの関連性を持っていると判断できるため、Web ページの組が登録されている利用者数を重みとすることで、上記の重み付きグラフを構築することができる。しかし、利用者の複数の興味を有しており、その興味に基づいて Web ページを登録することも考えられるため、そのような不要な繋がりを削除する必要がある。本論文では、重み付きグラフより定義される近傍情報を保存して低次元空間に写像することにより、上記の問題を解決する。これにより、データ全体から見た関連性の強い Web ページからなる近傍を抽出する。得られた Web ページの重み付きグラフを用いて k-means により Web ページのクラスタリングを行う。

提案手法の性能を検討するため、ソーシャルブックマークデータを模した人工データと Buzzurl ソーシャルブックマークデータを用いた実験を行った。人工データによる実験から、提案手法は関連性のあるデータ間の繋がりを残し、関連性の少ないデータ間の繋がりを効率的に削除できることを確認した。また、実データを用いた実験により、利用者の評価に基づいたクラスタリングが実現できることを確認できた。

連絡先: 柳本 豪一，大阪府立大学，堺市中区学園町 1-1，072-254-9279，072-254-9909，hidekazu@cs.osakafu-u.ac.jp

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

2. 従来研究

ソーシャルブックマークデータを用いた研究は多く行われている。特に、サーチエンジンなどのコンテンツに基づいたランキングとは異なり、多くの利用者の評価に基づいたランキングが可能である点に着目した研究が多い。以下では特に、クラスタリングに関する研究について紹介する。

Ramage らは Latent Dirichlet Allocation を用いた Web ページのクラスタリング手法を提案している [Ramage09]。この手法では、タグと登録された Web ページから抽出したキーワードを用いて分類し、タグによるクラスタリングはコンテンツに基づいたクラスタリングを異なる特徴があると述べている。また、Begeleman らは spectral bisection algorithm を用いてタグのクラスタリングを行っている [Begeleman06]。まず、タグの共起頻度をもとにグラフを作成し、そのグラフを分割することでタグのクラスタリングを行っている。Grahl らはタグの共起情報をもちいて、k-means を階層的に使い、タグを概念ごとにまとめている [Grahl06]。これらの手法は、共起情報をもとにクラスタリングを行っており、提案手法と似たものである。しかし、提案手法では Web ページ間の関連性に着目してノイズ削減を行い、クラスタリングを実現する。したがって、従来手法とはアプローチが異なるものである。

ノイズ削減手法として、提案手法関連のある手法として Laplacian eigenmap[Belkin02]、ISOMAP[Tanenbaum00]、Locally Linear Embedding[Roweis00] などがある。これらの手法は、高次元空間で表現されたデータの近傍情報を保存したまま、低次元空間にデータを写像することを実現する手法である。提案手法におけるノイズ削減手法は上記の手法を参考にしたものであるが、低次元化が目的ではなくノイズ削減が目的であり、これらの手法の目的とは異なるものである。

3. 提案手法

提案手法では、ユーザが登録している Web ページには関連性があると仮定し、Web ページの共起関係に着目したクラスタリング手法を提案する。しかし、利用者は複数の興味に基づいて Web ページの選別を行い、ソーシャルブックマークサー

ビスに登録を行う。このため、実際には登録した Web ページは必ずしも同一の興味によって判断されて登録されたわけではなく、Web ページ間では関連性が少ないものが同一利用者により登録されている場合がある。このような共起関係に着目した場合に発生する、見せかけの関連性を残したままクラスタリングを行うと精度の低下を招くと考えられる。以下では、このような Web ページ間の関連性をノイズと呼び、このノイズを削減することを考える。ノイズを低減する方法としては、主成分分析などのデータの分散に基づいた方法がある。しかし、提案手法ではデータの近傍をできるだけ維持した形でデータを低次元化で表現することにより、ノイズの削減を目指す。

まず、ソーシャルブックマークデータを重み付きグラフとして表現する。同一利用者に登録された Web ページの組は関連性があるとし、その Web ページの組が現れている利用者数を重みとして、重み付きグラフを作成する。ここで、上記のノイズに該当する繋がりは関連した Web ページ間で重みが小さいノードの組と見なすことができる。したがって、ノイズを低減することは、上記のような繋がりを削除すればよい。

今、Web ページ w_i と w_j 間の重みを K_{ij} とすると、上記のグラフは行列 $\mathbf{K} = K_{ij}$ で表現される。提案手法では、ノイズ削減は行列 $\mathbf{K}$ を用いて、Web ページ間の近傍を保存したまま、低次元空間への写像を決める処理となる。今、低次元空間のある軸上での Web ページ w_i の値を β_i とすると、Web ページ w_i と w_j が近い場合、β_i と β_j も近い値となることが上記の条件を満たすこととなる。これをユークリッド距離を用いて表現すると以下のようになる。

$$\min_{\beta} \sum_{i,j} K_{ij}(\beta_i - \beta_j)^2 \tag{1}$$

ここで、β は β_i からなるベクトルである。さらに式 (1) を書き換えると以下の様になる。

$$\sum_{i,j} K_{ij}(\beta_i - \beta_j)^2 = 2\beta^T \mathbf{P} \beta \tag{2}$$

ここで、行列 $\mathbf{P}$ は $\mathbf{P} = \mathbf{K} - \mathbf{\Lambda}$ と表現される。ただし、$\mathbf{\Lambda} = (\sum_i K_{1i}, \sum_i K_{2i}, \cdots, \sum_i K_{Ni})$ である。

ここで、Web ページ間の距離を保存した低次元空間での位置を決定する問題は、$\beta^T \mathbf{P} \beta$ を最小化するような β を決定する問題となる。ここで、以下の制約条件を導入する。

$$\beta^T \mathbf{\Lambda} \beta = 1 \tag{3}$$

上記の制約条件を用いて、ラグランジュの未定乗数法を用いて、目的の β を求める。

$$\mathrm{L} = \beta^T \mathbf{P} \beta - \lambda(\beta^T \mathbf{\Lambda} \beta) \tag{4}$$

上式を β で微分して 0 とおくと、以下の結果が得られる。

$$\mathbf{P}\beta = \lambda \mathbf{\Lambda} \beta \tag{5}$$

これは一般固有値問題であり、低次元空間でのデータの位置は一般固有値問題の固有ベクトルにより得られる。

上記の処理により、ノイズ削減の方法が一般固有値問題と一致したため、式 (1) を満たす β は最小固有値に対応する固有ベクトルから順番に選んでいけば良いこととなる。しかし、この一般固有値問題における最小固有値は 0 であり、それに対応する固有ベクトルとして、$\beta \propto \mathbf{1}$ が自明な解として存在する。この固有ベクトルは、すべて Web ページが同じ位置に存在することを表しており意味がない。したがって、自明な解である固有ベクトルは用いずに、最小固有値を除いて小さな固有値に対応する固有ベクトルを用いて処理を行う。

以上により、低次元空間での各軸に対する Web ページの重みが求められ、これらを用いて低次元空間での重み行列 $\mathbf{K}'$ を計算する。今、自明な解を除く k 個の固有ベクトルを並べた行列を $\mathbf{B}_k$ とする。この $\mathbf{B}_k$ を用いて $\mathbf{K}'$ を求めると以下となる。

$$\mathbf{K}' = \mathbf{\Lambda} \mathbf{B}_k (\mathbf{I} - \mathbf{D}_k) \mathbf{B}_k^T \tag{6}$$

ここで、$\mathbf{D}_k$ は対角成分に k 個の固有ベクトルに対応する固有値 λ_i を並べた行列である。以上の行列 $\mathbf{K}'$ は Web ページ間の強い関連性を残すことで得られた新しい重み付きグラフであり、弱い関連性は削除されノイズの低減が実現できている。

最後に、ノイズを削減した重み付きグラフ $\mathbf{K}'$ を用いた Web クラスタリングについて述べる。本論文ではクラスタリングには k-means を用いる。行列 $\mathbf{K}'$ は Web ページ間の類似度を表した行列であり、各 Web ページを特徴量で表現したベクトルとはなっていない。このため、従来の k-means を用いることはできず、カーネル k-means を用いることとする。以下では、カーネル k-maens の動作について説明する。ます、Web ページ w_i をベクトルで表現したものを $\phi(w_i)$ と表現する。ここで、i 番目のクラスタ N_i の重心を μ_i とすると、μ_i は以下のように表される。

$$\mu_i = \frac{1}{|N_i|} \sum_{w_i \in N_i} \phi(w_i) \tag{7}$$

次に Web ページ w_i がどのクラスタに属するか計算するため、$||\phi(w_i) - \mu_i||^2$ を求める。

$$\begin{aligned} ||\phi(w_i) - \mu_j||^2 &= \mathrm{k}(\phi(w_i), \phi(w_i)) \\ &- \frac{2}{N_i} \sum_{w_k \in N_j} \mathrm{k}(\phi(w_i), \phi(w_k)) \\ &- \frac{1}{|N_i|^2} \sum_{w_k \in N_j} \sum_{w_l \in N_j} \phi(w_k), \phi(w_l)) \end{aligned} \tag{8}$$

ここで、$\mathrm{k}(\phi(w_i), \phi(w_j)) = K'_{ij}$ とすると、行列 K' を用いた k-means が実現できる。

4. 実験

以下では 2 つの実験により提案手法の有効性を確認する。まず、人工的に作成したデータを用いて、提案手法が不要な関連性を削除した重み付きグラフを作成できることを示す。次に、Buzzurl ソーシャルブックマークデータを用いた web クラスタリングを行い、クラスタリング結果について検討を行う。

4.1 人工データ

ソーシャルブックマークデータを模倣した人工データを作成し、このデータを用いて提案手法の有効性を確認する。まず、70 名の利用者、100 件の Web ページが登録されているソーシャルブックマークデータを考える。ソーシャルブックマークデータは登録されているか、登録されていないかの 2 値データであるため、要素として 0 か 1 を持つ行列として表現される。また、各 Web ページ特定の話題を扱っていると見なして、ここで作成する人工データでは 5 つの話題が存在し、それぞれの話題を扱った Web ページが 10 件あるとする。したがって、

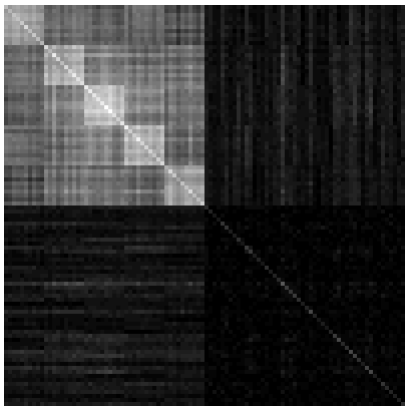

図 1: 人工データのグラフ表現

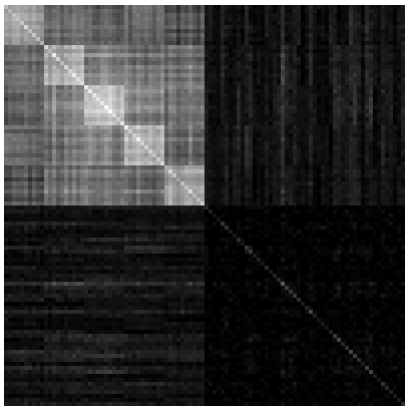

図 2: 主成分分析を用いた次元圧縮結果

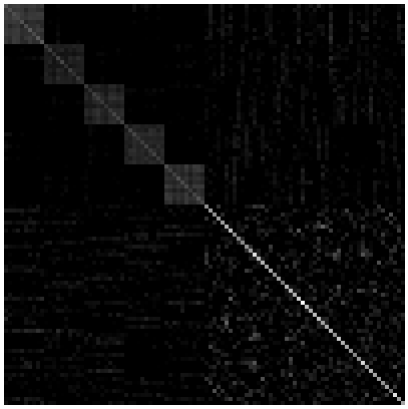

図 3: 提案手法を用いた次元圧縮結果

100 件の Web ページのうち、50 件が特定の話題を扱ったものであり、残りの Web ページは特定の話題を扱っていないものとして扱う。

一方、利用者は個人の興味に基づいて Web ページの登録を行う。本データでは、利用者は上記の 5 つの話題のうちどれかに興味があるとし、興味のある話題を扱う Web ページは高い確率で登録するものとする。ここで、利用者は複数の興味を持つことを許している。また、興味のない話題を扱う Web ページや特定の話題を扱っていない Web ページに関しても低い確率であるが、登録を行うものとする。以上の条件の下で、利用者の Web ページの登録状況を表した行列 $\mathbf{S}$ を作成する。行列 $\mathbf{S}$ は 70×100 の行列である。

次に、行列 $\mathbf{S}$ を用いて、Web ページ間の重み付きグラフを表す行列 $\mathbf{K}$ を作成する。行列 $\mathbf{K}$ は 100×100 の対称行列であり、以下により求める。

$$\mathbf{K} = \mathbf{S}^T\mathbf{S} \tag{9}$$

図 1 に作成された行列 $\mathbf{K}$ を示す。縦と横は Web ページを表し、それぞれの点は Web ページ間の共起回数が多いものほど白く表示している。ただし、Web ページの順番は同一の話題を扱ったものが順番になるように並び替えている。この図から、同一の話題を扱っている Web ページの部分は白くなっていることが分かる。また、利用者が複数の話題を有するため、話題を扱った Web ページ間では頻繁に共起していることが分かる。一方、特定の話題を扱っていない Web ページはランダムに登録されるため、他の Web ページとはあまり共起していないことも分かる。

このデータに対して、提案手法と主成分分析を適用し、50 個の固有ベクトルを用いて行列 $\mathbf{K}'$ を作成する。主成分分析による次元圧縮について簡単に説明を行う。まず、行列 $\mathbf{K}$ を固有値分解する。

$$\mathbf{K} = \mathbf{V}^T\mathbf{D}\mathbf{V} \tag{10}$$

ここで、$\mathbf{V}$ は固有ベクトルからなる行列、$\mathbf{D}$ は固有値を対角成分として持つ行列である。ここから、最大固有値から k 個の大きな固有値を取り出すことで、次元圧縮を行う。今、k 個の固有値からなる行列を $\mathbf{D}_k$、k 個の固有ベクトルからなる行列を $\mathbf{V}_k$ とすると、行列 $\mathbf{K}'$ は以下のように表される。

$$\mathbf{K}' = \mathbf{V}_k^T\mathbf{D}_k\mathbf{V}_k \tag{11}$$

主成分分析により作成された結果を図 2、提案手法により作成された結果を図 3 に示す。主成分分析により作成された行列 $\mathbf{K}'$ は本来の $\mathbf{K}$ を近似していると言えるが、異なった話題間の関連性が残ったままであり、十分にノイズが削除されているとは言いにくい。一方、提案手法による手法では同一の話題を含む Web ページ間の重みを残し、異なる話題を含む Web ページ間の重みが小さなものとなっていることが分かる。これは、本論文の目的であるノイズの削除が行えていると言える。

4.2 実データ

Buzzurl ソーシャルブックマークデータを用いて Web ページのクラスタリングを行い、提案手法の有効性を確認する。Buzzurl ソーシャルブックマークデータは 25,597 名の利用者、864,574 件の Web ページからなるデータであり、利用者と Web ページの組からなるデータは 1,626,869 件である。このデータは 2005 年 10 月から 2008 年 10 月までに登録されたものである。

本実験ではある程度の利用者が評価した Web ページを対象としてクラスタリングを行うことを目的とするため、以下の条件でデータの選別を行った。

1. 10 件以上の Web ページを登録している利用者
2. 20 名の利用者から登録されている Web ページ

本手法は同時に登録されている Web ページに注目するため、登録数が少ない利用者からはそのような Web ページの組が発

表 1: クラスタリング結果例 1

http://japan.cnet.com/
http://kakaku.com/
http://www.youtube.com/
http://www.checkpad.jp/
http://ja.wikipedia.org/wiki/%E3%83%A1%E3%82%A4%E3%83%B3%E3%83%9A%E3%83%BC%E3%82%B8

表 2: クラスタリング結果例 2

http://www.youtube.com/watch?v=P6uFXSE3ARM
http://www.youtube.com/watch?v=PRma29nAp9s
http://www.youtube.com/watch?v=GDk8qxztJLQ
http://www.youtube.com/watch?v=PC4HhI0bK4c
http://www.youtube.com/watch?v=bq5jkIkhndM

見できないため、Web ページの登録数を閾値として用いた。また、Web ページについても Web ページ間の関連性を同時に登録している利用者数で表現するため、多くの利用者から評価される必要があり、一定数の利用者から登録されているものに限定した。この条件より、2,256 名の利用者、2,707 件の Web ページが実験対象として選ばれた。

次に、上記から選ばれた利用者と Web ページを対象にソーシャルブックマークデータから行列 $\mathbf{S}$ を作成する。この行列は、人工データでの実験で説明を行った行列と同じものである。行列 $\mathbf{S}$ は $2,256 \times 2,707$ の行列であり、以下で定義する。

$$S_{ij} = \begin{cases} 1 & (\text{利用者 } u_i \text{が Web ページ } w_i \text{を登録している時}) \\ 0 & (\text{その他}) \end{cases} \tag{12}$$

そして、行列 $\mathbf{S}$ を用いて行列 $\mathbf{K}$ を以下のように求める。

$$\mathbf{K} = \mathbf{S}^T\mathbf{S} \tag{13}$$

この行列は $2,707 \times 2,707$ の対称行列である。

次に、提案手法を用いて行列 $\mathbf{K}$ を低次元化を行い、得られた行列 $\mathbf{K}'$ を用いてカーネル k-means により Web ページのクラスタリングを行う。クラスタリング結果の一例を表 1 と表 2 に示す。

表 1 はサイトのトップページが集まったクラスタであり、利用者の情報源となるサイトを表している。このようなサイトは Yahoo!などのポータルサイトと異なり、利用者の興味によりまとめられた有益なサイトであり、通常は同じくラスタとはならないと思われる。一方、表 2 は Youtube のサイトが集まっている。現在公開されていないものもあるが、利用者が付けたタグから判断すると、すべて猫に関する動画である。本手法では、タグや動画のタイトルなどのコンテンツに関する情報を扱っていない。しかし、利用者の登録状況から内容が似た Web サイトを同一のクラスタにまとめることができることを表している。以上より、利用者の登録情報を用いて、Web ページをクラスタリングすることが有効であることが確認できた。

一方、 得られたクラスタには関連性が明確には分からないものが多く含まれていた。この理由として、(1) 気づかなかった関連性によりまとまっている、(2) クラスタ数が適切でなく不要なクラスタが作られた、というものが考えられる。Web ページ間の関連性については、多面的に検討する必要があるため、クラスタに関する詳細な検討を行う必要がある。一方、クラスタ数についても適切なクラスタ数を決定する手法を検討する必要があり、これも今後の課題であると考えている。

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手法を提案した。本手法では、まず 2 つの Web ページが同時に登録されている頻度に着目し、Web ページの重み付きグラフを作成した。そして、得られたグラフから Web ページ間の近傍を考え、その近傍情報をできるだけ保存した形で低次元化することで、不要な関連性を削除することを実現した。

人工データにより、提案手法では関連性のある Web ページの繋がりを残し、関連性のない Web ページの繋がりを効率的に削除できることを確認した。また、Buzzurl ソーシャルブックマークデータを用いた実験により、関連性の強い Web ページを同一クラスタにまとめることができることを確認した。また、得られたクラスタが利用者の評価に基づいたものであることも確認した。

しかし、以下の課題がまだ残っているため、この課題に今後取り組んでいく予定である。

1. 適切なクラスタ数の決定方法
2. クラスタ内の Web ページの関連性の検討

最後に本研究に対してソーシャルブックマーク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

[Ramage09] Ramage D., Heymann P., Manning C. D., and Garcia-Molina H.: Clustering the Tagged Web, Proceedings of the Second ACM International Conference on Web Search and Data Mining, pp.54-63, (2009)

[Begeleman06] Begeleman G., Keller P., and Smadja F.: Automated Tag Clustering: Improving search and exploration in the tag space, Collaborative Web Tagging Workshop, 15th WWW Conference Edinburgh, (2006)

[Grahl06] Grahl M., Hotho A., and Stumme G.: Conceptual Clustering of Social Bookmarking Sites, 7th International Conference on Knowledge Management, Austria, (2007)

[Belkin02] Belkin M. and Niyogi P.: Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering, Advances in Neural Processing Systems, pp.585-591, (2002)

[Tanenbaum00] Tanenbaum J. B., de Silva V., and Langford C.: A Gloval Framework for Nonlinear Dimensionality Reduction, Science, Vol.209, pp.2319-2323, (2000)

[Roweis00] Roweis S. and Saul L.: Nonlinear Dimensionality Reduction by Locally Linear Embedding, Science, Vol.209, pp.2323-2326, (2000)