2E2-4

タグの語彙の違いを考慮したソーシャルブックマークデータのグラフ表現

Graph Representation of Social Bookmarking Data Regarding Difference between Users' Tag Vocabulary

柳本 豪一[*1] 吉岡 理文[*1]
Hidekazu Yanagimoto Michifumi Yoshioka

[*1]大阪府立大学
Osaka Prefecture University

These days social media services are widespread and make social networks in the Internet. Community detection and knowledge discovery are realized analyzing the networks. A social bookmarking service is one of the social media services and is paid attention to as a new information resource. In the social bookmarking service users shared their bookmarks with other users. Our aim is to make web pages' graph from social bookmarking data and to analyze it. We especially propose a method to construct the graph from the social bookmarking data and evaluate the proposed method using real social bookmarking data. Concretely similarities between web pages are defined using tag vocabulary for each user and the similarities are integrated to construct all web pages' graph. To integrate the similarities we use Bayes estimation.

1. はじめに

近年，ソーシャルメディアが新しい情報源として注目を浴びている．従来のコンテンツの内容に基づいたリンクから構成されるネットワークとは異なり，ソーシャルメディアは人間関係に基づいている．ソーシャルメディアの代表例としては，Facebook[*1]，mixi[*2]などのソーシャルネットワークサービス，del.icio.us[*3]，はてな[*4]，Buzzurl[*5]などのソーシャルブックマークサービスがある．ソーシャルメディアはネットワークとして表現される．このネットワーク構造に基づいてノードのクラスタリング [1] が注目されている．例えば，コミュニティ解析 [2] を行うことで，関連性の高いノードをグループ化することができる．さらに，このクラスタリングを利用することで，関連のある情報を見つけることができる．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際，ノードは Web ページとなる．しかし，人間関係とは異なり，明示的に Web ページ間の関連性を表す情報はソーシャルブックマークサービスでは用意されていない．このため，Web ページ間の関連性を定義する必要がある．

本論文では，ソーシャルブックマークサービスを対象として，ユーザが登録した Web ページをネットワークで表現する手法を提案する．具体的には，ソーシャルブックマークサービスでユーザが付加したタグを用いて Web ページ間の類似度を定義する．このとき，タグの語彙がユーザごとに異なる点に注意し，ネットワークの構築を行う必要がある．例えば，ユーザ間で同一タグが異なる視点から利用されている可能性があるため，Web ページ間に不要な類似度が定義されることを避けるなどである．これを実現するため，ユーザごとに Web ページの類似度を定義し，ユーザごとの類似度を統合することで上記の問題を回避する．最後に，Buzzurl ソーシャルブックマークサービスのデータを用いて評価実験を行い，同一ユーザに登録された Web ページ群は関連性があるとした手法に比べ，提案手法が不要なリンクを生成しないことを確認した．

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://www.facebook.com/
∗2 http://mixi.co.jp/
∗3 http://www.delicious.com/
∗4 http://www.hatena.ne.jp/
∗5 http://buzzurl.jp/

2. 従来研究

ユーザ間の関連性に関する情報は，既にソーシャルメディアのデータに含まれており，一般的にユーザ間のリンクはそれを用いて定義される．例えば，文献 [3] では，ユーザ間でのメールの送受信の回数をもとにユーザ間の関連性を定義している．文献 [4] では，E-mail，Instant Message など複数のコミュニケーションツールの利用状況を把握してユーザ間の関連性を定義している．文献 [5] では，社会ネットワーク [6] の研究で知られている Homophily[7] に着目し，社会ネットワークの時間的な変化について検討を行っている．これより，ユーザの属性の類似性を用いてネットワークを構成することも可能である．文献 [8] では，Twitter の Follow 情報をもとにネットワークを構築している．ユーザの関係が Follow とうい形で明示されているので，ネットワークを構成することが容易である．

Web ページの類似度の定義に関する関連研究について紹介する．文献 [9] では，Web ページの共起に基づいて類似度が定義されている．ブックマークフォルダごとに Web ページが管理されていると想定し，同一フォルダに含まれている Web ページは関連性があるとしている．ソーシャルブックマークサービスでは，登録 Web ページはタグの Folksonomy[10] を用いて管理されているので，上記の手法をそのまま応用することはできない．文献 [11] は，タグを共有する Web ページ間には関連性があるとしてネットワークを構築している．そして，この実験よりタグの多義性のため，異なった内容の Web ページに関連性が作成されることを指摘している．

3. 提案手法

提案手法は，(1) ユーザごとにタグを用いて Web ページの類似度を定義する，(2) 個々のユーザごとに定義された類似度を統合する，の 2 段階から構成されている．特に，統合する際にはベイズ推定にもとづいて，統合後の類似度を求める．これにより，グラフ作成者の事前知識を反映したり，個々のユーザの評価を反映したネットワークの構成が容易に行える．

3.1 ユーザごとの Web ページの類似度の定義

ソーシャルブックマークはユーザが興味のある情報 (Web ページなど) を登録し，公開するサービスである．登録されたデータにはタグと呼ばれるキーワードを付けることが可能である．このタグはユーザが自由に設定できるものであり，語彙は制限されていない．このため，同じ内容を表していると思われる Web ページに異なるタグが用いられていたり，同じタグが異なった意味で利用されることが起こる．例えば，文献 [11] では"sf"という単語が"science fiction"と"San Francisco"の 2 つの意味を持っており，ソーシャルブックマークサービスで用いられていると述べられている．これはタグの多義性に起因する例である．これを回避するためには，タグを用いて Web ページ間に類似度を計算する際に対処が必要である．

今，ユーザごとにタグの集合を考える．タグは登録された Web ページをユーザが管理するために利用するため，ユーザごとにタグの語彙を考えることで上記の問題を回避できると考えられる．なぜなら，同じタグを異なる意味で用いた場合，異なる web ページが関連付けられ，自分の登録 Web ページの管理が困難になるからである．したがって，上記の設定では，タグの多義性の影響は小さいと考えられる．

本手法では，タグを用いた Web ページの類似度の定義には Jaccard 係数を用いる．Jaccard 係数は集合間の類似度を定義する手法であり，Web ページに付加されたタグの集合を比較することで，Web ページの類似度を計算できる．

$$\mathrm{sim}^k_{i,j} = \frac{|T_{ki} \cap T_{kj}|}{|T_{ki} \cup T_{kj}|} \tag{1}$$

ここで，T_{ki} はユーザ u_k が Web ページ w_i に付加したタグの集合，$|\cdot|$ は集合の要素数を表す．共通したタグが多く共有されている Web ページ間の類似度は大きくなる．

本手法では，タグの表記の揺れ等を自然言語処理などにより一切補正せず，ユーザが登録したものを利用する．タグは単語のみの場合が多く，表記もユーザごとに異なっているため，適切な修正で表記の揺れを解決することは難しいと判断したためである．類語辞書等を用いて表記の揺れ等を解決できるとも考えられるが，辞書に記載されていないものもタグとして多く使われているため，有効とは言い難い．さらに，ユーザによってはわずかな表記の違いによってタグの意味の違いを表している場合もあり，これを補正することは好ましくない．以上の点を考えて，表記の揺れ等を補正しないという方針を採用した．

3.2 ユーザ間の Web ページの類似度の統合

ユーザが閲覧できる Web ページは全 Web ページの一部であり，また興味のある Web ページの一部しか閲覧できない．したがって，ユーザごとに定義された Web ページの類似度を統合し，全 Web ページの類似度を求めてグラフを作成する必要がある．このとき，異なるユーザで同じ Web ページ間の類似度が異なる可能性があるので，異なった類似度が得られた場合に類似度を統合する手法を提案する．

今，ユーザごとに定義された Web ページ間の類似度について検討する．もし，真の Web ページ間の類似度が存在すると仮定すると，ユーザごとに定義された Web ページ間の類似度のばらつきは観測値のばらつきと見なせる．

$$\mathrm{sim}^k_{i,j} = \mathrm{sim}^*_{i,j} + \epsilon_k \tag{2}$$

今，$\mathrm{sim}^k_{i,j}$ はユーザ u_k により定義された Web ページ w_i と w_j の類似度である．また，$\mathrm{sim}^*_{i,j}$ は Web ページ w_i と w_j の類似度の真値，ϵ_k は観測による誤差であり，平均 0，分散 σ_k^2 の正規分布 $\mathrm{N}(0, \sigma_k^2)$ に従う．$\mathrm{sim}^k_{i,j}$ は以下のように定義される．

$$\Pr(\mathrm{sim}^k_{i,j}|\mathrm{sim}^*_{i,j}) = \mathrm{N}(\mathrm{sim}^*_{i,j}, \sigma_k^2) \tag{3}$$

次に，$\mathrm{sim}^*_{i,j}$ について考える．あらかじめ Web ページ間の類似度が他手法（コンテンツの内容の類似性など）で推定されている場合，そのような知識を反映できれば柔軟に真値の推定を行える．これに対処するため，$\mathrm{sim}^*_{i,j}$ の事前分布を考える．

$$\Pr(\mathrm{sim}^*_{i,j}) = \mathrm{N}(\mu, \sigma_0^2) \tag{4}$$

複数の観測値が得られたときに真の類似度の推定を行う．推定には，ベイズ推定を用いる．今，n 名のユーザから Web ページ w_i と w_j の類似度が定義されている場合を考える．事後分布 $\Pr(\mathrm{sim}^*_{i,j}|\mathrm{sim}^1_{i,j}, \cdots, \mathrm{sim}^n_{i,j})$ は，以下のように定義される．

$$\begin{aligned}
& \Pr(\mathrm{sim}^*_{i,j}|\mathrm{sim}^1_{i,J}, \cdots, \mathrm{sim}^n_{i,j}) \\
= \; & \frac{\prod_k \Pr(\mathrm{sim}^k_{i,j}|\mathrm{sim}^*_{i,j})\Pr(\mathrm{sim}^*_{i,j})}{\sum_{\mathrm{sim}^*_{i,j}} \prod_k \Pr(\mathrm{sim}^k_{i,j}|\mathrm{sim}^*_{i,j})\Pr(\mathrm{sim}^*_{i,j})} \\
\propto \; & \prod_k \Pr(\mathrm{sim}^k_{i,j}|\mathrm{sim}^*_{i,j})\Pr(\mathrm{sim}^*_{i,j}) \\
= \; & \mathrm{N}(\mu'|\sigma'^2)
\end{aligned} \tag{5}$$

ここで，μ' と σ'^2 は以下のように求められる．

$$\mu' = \frac{\frac{\mu}{\sigma_0^2} + \sum_k \frac{\mathrm{sim}^k_{i,j}}{\sigma_k^2}}{\frac{1}{\sigma_0^2} + \sum_k \frac{1}{\sigma_k^2}} \tag{6}$$

$$\frac{1}{\sigma'^2} = \frac{1}{\sigma_o^2} + \sum_k \frac{1}{\sigma_k^2} \tag{7}$$

以上の事後分布から類似度の推定値 $\mathrm{sim}_{i,j}$ を求める．今回は事後分布のモードを用いると，平均値となる．

$$\mathrm{sim}_{i,j} = \frac{\frac{\mu}{\sigma_0^2} + \sum_k \frac{\mathrm{sim}^k_{i,j}}{\sigma_k^2}}{\frac{1}{\sigma_0^2} + \sum_k \frac{1}{\sigma_k^2}} \tag{8}$$

分散 σ_k^2 は推定値の信頼を表す指標として利用することができる．信頼できるユーザの評価の場合には分散を小さくし，信頼できないユーザの場合は分散を大きくすることで，ユーザで定義された類似度の影響をコントロールすることが可能である．例えば，スパムユーザに対しては分散を大きくするなどの対応をとることができる．

4. 実験と考察

本章では実際のソーシャルブックマークデータからグラフを構成し，そのグラフを検討する．以下では，実験で用いたデータの詳細と作成したグラフにおいてどのような Web ページ間に類似度が定義されているかについて説明する．

4.1 実験データ

今回の評価実験には Buzzurl ソーシャルブックマークサービスのデータを用いた．このデータは 2005 年 10 月から 2008 年 10 月までに登録された 2008 年 10 月時点での全データである．データの構成を表 1 に示す．1 つのソーシャルブックマークデータは「ユーザ」,「Web ページ」,「タグ集合」から構成

表 1: Buzzurl ソーシャルブックマークサービスのデータの構成

registered data	1,626,869
unique URLs	864,574
unique users	25,597
unique tags	352,016

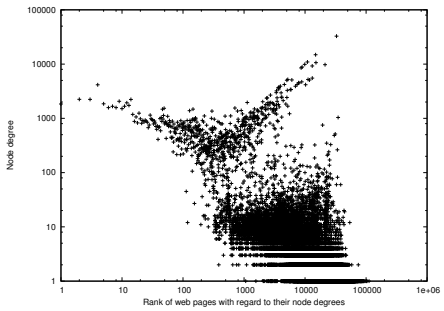

図 1: ユーザに同時に登録された頻度に基づいたグラフの次数分布

され，表中の”registered data”で表されている．評価実験では約 86 万件のすべての Web ページを用いた．

このデータに付加されたタグに付いて調べると，ソーシャルブックマークデータのうち約 4 分の 1 にタグが付加されていなかった．したがって，本手法ではタグを用いて Web ページの類似度を定義するため，タグを含んでいないソーシャルブックマークデータを利用することはできない．

4.2 ソーシャルブックマークデータからのグラフ作成

まず，比較手法について説明する．比較手法としては，ユーザが登録している Web ページはすべて関係があると見なして，グラフを作成する手法である．つまり，ユーザにおける Web ページの共起に着目した手法である．この手法では，ユーザが複数の興味を有している場合，不要な Web ページ間の関連性を生成する可能性がある．この手法では，類似度は 2 つの Web ページ対を登録したユーザ数で定義する．

比較手法で得られた各ノード (Web ページ) の次数の分布を図 1 に示す．この図から，この手法で作成されたグラフはべき乗分布をなさず，次数が高いノード数の多いことが分かる．

次に，同様に比較手法を用いて，2 名以上のユーザから同時に登録されている Web ページのみを用いるという閾値を導入した場合を図 2 に示す．一度しか同時に登録されていない Web ページ対を削除しただけで，図 1 に比べて，べき乗分布が明確になっている．閾値を導入することは，頻繁に同時に登録されている Web ページ対のみを扱うことであり，関連性の強いもののみを扱っていることである．

次に提案手法の結果について述べる．提案手法では μ，σ_0^2，σ_k^2 を決定する必要である．本実験では，個々のユーザの分散は同じであると仮定し，$\mu = 0.5$，$\sigma_0^2 = 10$，$\sigma_k^2 = 1$ とした．提案手法で得られた各ノードの次数分布を図 3 に示す．この結果では閾値を導入していないが，べき乗分布が見られる．

表 2 に各手法から得られたリンク数を示す．この結果より，提案手法を用いた方が約 10 分の 1 のリンク数となり，不要なリンクを生成していないことが分かる．

次に，”http://www.google.co.jp/”に注目して作成されたグラフの特徴について述べる．表 3 に各手法から得

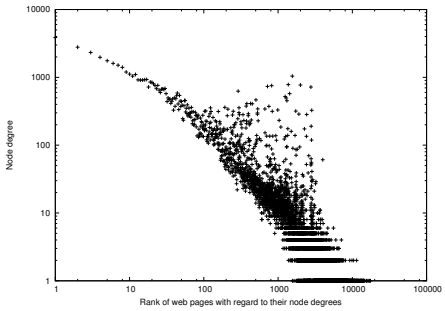

図 2: 2 名以上のユーザに同時に登録された頻度に基づいたグラフの次数分布

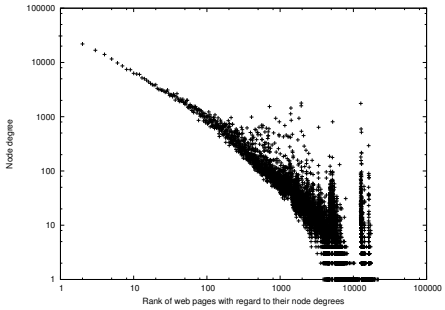

図 3: 提案手法により作成されたグラフの次数分布

表 2: 各手法から得られたリンク数

Method	Links
Cooccurrence	2,300,343,319
Jaccard	284,561,877

表 3: ”http://www.google.co.jp/”のノードの次数

Method	Degree
Cooccurrence	15,565
Jaccard	1,137

られた”http://www.google.co.jp/”のノードの次数を表す．これより，提案手法は次数が少ないことが分かる．次に，”http://www.google.co.jp/”と関連づけられた Web ページの上位 6 件を表 4 に示す．提案手法で上位の Web ページを選ぶ際には，1 名だけのユーザで類似度が推定されている場合を除外した．なぜなら，観測値が少なく，推定値の信頼性が低いからである．結果は提案手法ではサーチエンジンや検索関連の Web ページが上位に位置している．

4.3 考察

まず，図 1，図 2，図 3 を用いて，各手法より得られたグラフの特性について検討する．ソーシャルブックマークサービスでは，人気のある話題を扱った Web ページは多く登録され，関連した Web ページも多く登録されると考えられる．また，少数のユーザしか興味を持たない話題を含んだ Web ページも同様に登録されるが，関連 Web ページも少ししか登録されないと考えられる．しかし，個人の嗜好を反映しているため，多

表 4: ”http://www.google.co.jp/”と関連づけられた Web ページのリスト

Cooccurrence	Jaccard
http://www.yahoo.co.jp/	http://www.excite.co.jp/
http://www.asahi.com/	http://www.yahoo.co.jp/
http://www.rakten.co.jp/	http://www.hatena.ne.jp/
http://www.livedoor.com/	http://find.2ch.net/
http://jp.msn.com/	http://ask.jp/
http://www.vector.co.jp/	http://jp.msn.com/

くの種類が含まれると思われる．このような特徴は，ノードの次数の分布としてはべき乗分布として表現されるものである．上記の仮定の下で分布のグラフを見ると，提案手法はべき乗分布を表しており，適切なグラフが構成されていると思われる．一方，比較手法では Web ページ対の出現頻度に閾値を導入することにより，べき乗分布が現れた．これより，ユーザが登録したすべての Web ページ対に関連性があると見なすことは，過剰に関連性を生成することを表している．

次に，表 2 と表 3 を用いて，生成されたリンク数について検討する．タグはコンテンツの内容およびユーザの判断に基づいて付加されるものであり，関連のある Web ページには同じタグが用いられると考えられ，提案手法は不要なリンク生成を防いでいる．このため，比較手法に比べて約 10 分の 1 のリンク数になったと思われる．特定の Web ページの次数を調べた場合も表 3 に示されるように，同様の傾向が見られた．

最後に表 4 の検討を行う．”http://www.google.co.jp/”との類似度が高い上位 6 件の Web ページのリストを見ると，提案手法の方が関連性の高い Web ページが選択できている．比較手法では登録頻度が高い Web ページとの類似度が大きくなる傾向があり，ニュースサイトなど関連が薄いと思われる Web ページが上位に入っている．なぜなら，比較手法では数多く登録されている Web ページは多くの Web ページと同時に登録される可能性が高いからである．一方，提案手法ではユーザが付加したタグを用いて類似度を計算するため，登録数に関係なく関連性のある Web ページの類似度が高くなる．このため，”http://www.excite.co.jp/”がトップになったと思われる．

5. おわりに

ソーシャルブックマークサービスを対象として，登録された Web ページ集合をグラフとして表現する手法を提案した．具体的には，ユーザごとにタグを用いて Web ページの類似度を計算し，全ユーザから得られた Web ページ間の類似度を統合することで，全登録 Web ページ間の類似度を求めた．統合による類似度推定にはベイズ推定を用いた．

Buzzurl ソーシャルブックマークデータを用いた評価実験を行い，ユーザに同時に登録された回数による類似度決定手法と比較を行った．提案手法は比較手法に比べて，約 10 分の 1 のリンクしか生成しないことが分かった．生成されたリンクの質については，”http://www.google.co.jp/”とリンクされた Web ページを調べることで，提案手法は関連性の強い Web ページに高い類似度が割り当てられていることが分かった．以上より，提案手法はソーシャルブックマークデータから質の高いグラフを作成することができることが分かった．

今後は，ソーシャルブックマークデータから作成した Web ページのグラフを解析することで，Web ページのクラスタリングや知識発見を行う予定である．

謝辞

Buzzurl ソーシャルブックマークサービスのデータ提供に付いて EC ナビ株式会社に感謝いたします．

参考文献

[1] Shi, J. and Malik, J.: Normalized Cuts and Image Segmentation, PAMI, Vol.22, No.8, pp.888-905(1972)

[2] Tang, L. and Liu H.: Community Detection and Mining in Social Media, Synthesis Lectures on Data Mining Knowledge Discovery, Moragan & Claypool Publishers(2010)

[3] Choudhury, D., M., Mason, A., W., Hofman J., M. and Watts D., J.: Inferring Relevant Social Networks from Interpersonal Communication, Proc. of the 19th Intenational Conference on World Wide Web, pp.301–310(2010)

[4] Horowitz, D. and Kamvar S. D.: The Anatomy of a Large-Scale Social Search Engine, Proc. of the 19th International Conference on World Wide Web, pp.431–440(2010)

[5] Fond, T. L. and Neville, J.: Randomization Tests for Distinguishing Social Influence and Homophily Effects, Proc. of the 19th International Conference on World Wide Web, pp.601–610(2010)

[6] Wasserman, S. and Faust, K.: Social Network Analysis, Cambridge University Press(1994)

[7] McPherson, M. , Smith-Lovin, L. and Cook, J.: Birds of a feathre: Homophily in social networks, Annual Review of Sociology, Vol.27, pp.415–445(2001)

[8] Kwak, H., Lee, C., Park, H. and Moon, S.: What is Twitter, a Social Network or a News Media?, Proc. of the 19th International Conference on World Wide Web, pp.591–600(2010)

[9] Rucker, J. and Polanco M., J.: Siteseer: Personalized navigation for the web, Comm. ACM, Vol.40, No.3, pp.73–75(1997)

[10] Mathes, A.: Folsonomies – Cooperative Classifiction and Communication Through Shared Metadata, http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

[11] Yeung C., A., Gibbins, N. S. and Shadbolt, N.: Understanding the Semantics of Ambiguous Tags in Folksonomies, The International Workshops on Emergent Semantics and Ontology Evolution at ISWC/ASWC 2007(2007)