ソーシャルブックマークにおけるユーザのタグ付け傾向を加味した Web ページ推薦手法

情報数理応用研究

5209C012-2　岸端佑季
指導教員　後藤正幸

Web Page Recommendation Based on User's Tagging Tendency in Social Bookmark

KISHIBATA Yuuki

1　はじめに

近年, インターネット上に存在する Web ページは飛躍的に増加しており, 莫大な数の Web ページの中から, ユーザの興味を満たすページを自動的に発見してくれる推薦システムの重要性が増している. 一方, はてなブックマーク [1], del.icio.us[2] のように, 一つのサイト上で複数のユーザのブックマークを共有することができるソーシャルブックマーク (以下, SBM) と呼ばれるサービスが台頭している. ブックマークとは, ユーザがお気に入りと登録した Web ページのことである. SBM の特徴は, 自分のブックマークに対し, 特徴を表すキーワードとして, タグと呼ばれるメタデータを付与することができることである. ユーザにとって, タグとは自分のブックマークを読み直すための分類カテゴリとして機能しており, 独自の視点, 規則でタグを付与してそれらを管理している. タグやブックマークは, SBM 上に日々蓄積され続けているが, Web ページ推薦システムにとって重要なユーザの興味を表現する情報として活用できる可能性を秘めている [3].

このような背景から, SBM を活用した Web ページ推薦システムに注目が集まっている. これまでに, あるユーザが利用しているタグの利用履歴とある Web ページに付与されてきたタグの付与履歴を比較し, その一致度が高いときに Web ページを推薦するシステム [3], あるユーザが同じタグ 1 個を付与しているブックマーク集合同士の類似度を計算し, 類似度の高いブックマーク集合に含まれる Web ページを推薦するシステム [4] などが提案されている. 前者は比較する履歴間のタグの名称が一致しても, 異なるユーザが異なる意味合いでそのタグを使用している場合は推薦精度が低下する. 後者のシステムでは, ブックマーク集合に含まれる Web ページの内容は, 単一のトピックに関するものであることを前提としている. しかし, ユーザが興味を示すトピックには, 個々のユーザ毎に異なる階層的な構造があると仮定すると, トピックは, ユーザ毎に興味が異なるようなサブトピックに分割できると考えられる. 例えば, スポーツというトピックであれば, あるユーザは野球やサッカーというサブトピック, 別のユーザであればラグビーや水泳というサブトピックというように, スポーツというトピックを個々のユーザ毎に興味が異なるサブトピックに分けることができる. 後者のシステムでは, ユーザのサブトピックに対する興味を考慮しておらず, ブックマーク集合同士の類似性を測る際, 特定のサブトピックに関して興味が部分的に類似するブックマーク集合を考慮することで, さらに推薦精度の向上が期待できる.

本研究では, 後者のシステムを改良し, ユーザが興味を示すトピックには個々のユーザ毎に異なる階層的な構造があると仮定し, 特定のサブトピックに関して部分的に類似している興味を抽出することで, 推薦精度を向上させる手法を提案する. 実際の評価実験を通じ, 提案手法の有効性を示す.

2　準備

本研究で扱う用語および記号の定義を記述する. Web ページは単一の URL とし, ブックマークはユーザがお気に入り登録した Web ページとする. SBM を利用しているユーザを $u_i(i = 1, 2, \cdots, I)$, ユーザ u_i が利用している全てのタグ集合を $\mathcal{T}_i = \{t_1^{(i)}, t_2^{(i)}, \cdots, t_{J_i}^{(i)}\}$, SBM 内のユーザにブックマークとして登録されたことがある Web ページ集合を $\mathcal{P} = \{p_1, p_2, \cdots, p_W\}$, ユーザ u_i のブックマーク集合を $\mathcal{B}_i \in \mathcal{P}$, ユーザ u_i がタグ $t_j^{(i)}$ を付与した u_i のブックマークの集合を $\mathcal{B}_j^{(i)} \in \mathcal{P}$ とする.

3　従来研究

3.1　従来研究の位置づけ

丹羽ら [3] は, ユーザの興味を表現する情報としてタグの名称を利用しており, あるユーザが SBM で利用しているタグの利用履歴とある Web ページに付与されてきたタグの付与履歴を比較し, 履歴間のタグの名称の一致度が高いとき, ユーザに Web ページを推薦している. しかし, 同じタグを付与しても, 異なるユーザが同じ意味でそのタグを使用しているとは限らない [4]. 例として, 同じ「スポーツ」というタグに対して, あるユーザは「野球」, 他のユーザは「サッカー」と考え, タグを付与している可能性がある. このようなケースが SBM 内のタグには多く存在するため, 必ずしも精度が良くならないという問題がある [4].

次に, 丹羽らの推薦システムの問題点を改善した佐々木ら [4] による推薦システムに関して述べる. 佐々木ら [4] は, ブックマーク集合同士の類似性から Web ページを推薦するシステムを考案した. 佐々木らは, あるユーザ u_i がタグ $t_j^{(i)}$ を付与している u_i のブックマーク同士は内容が類似していることに着目し, これらのブックマーク集合 $\mathcal{B}_j^{(i)}$ を u_i のタグ $t_j^{(i)}$ に関するブックマーククラスタと定義している. ブックマーククラスタはユーザの興味を表現している. 図 1 は, ブックマーククラスタの例である. 例えば, あるユーザの「ビジネス」というタグに関するブックマーククラスタは, そのユーザが「ビジネス」に興味があるという情報を表している.

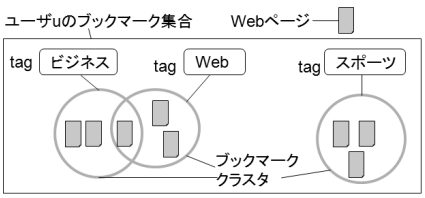

図 1. ブックマーククラスタ

したがってこの手法では, このブックマーククラスタを利用して Web ページを推薦する．推薦を受けるユーザのブックマーククラスタとそれ以外のユーザのブックマーククラスタの類似度を, ブックマーククラスタ間に共通する Web ページの数を基に算出する．そして, 類似度の高いブックマーククラスタに含まれる Web ページを推薦する．

3.2 ブックマーク集合の類似性による推薦システム

以下で, ブックマーク集合の類似性による推薦手法の詳細について説明する．

3.2.1 ブックマーククラスタ同士の類似度の算出方法

ブックマーククラスタ同士の類似度を，二項分布を利用した対数尤度比の概念を利用して算出する．ここで，ユーザ u_i のタグ $t_j^{(i)}$ に関するブックマーククラスタ $\mathcal{B}_j^{(i)}$ とユーザ $u_{i'}$ のタグ $t_{j'}^{(i')}$ に関するブックマーククラスタ $\mathcal{B}_{j'}^{(i')}$ との類似度を算出することを考える．どちらのユーザもお気に入り登録していて，一方のブックマーククラスタに含まれている Web ページの集合 $N(\mathcal{B}_j^{(i)}, \mathcal{B}_{j'}^{(i')})$ を

$$N(\mathcal{B}_j^{(i)}, \mathcal{B}_{j'}^{(i')}) = (\mathcal{B}_i \cap \mathcal{B}_{i'}) \cap (\mathcal{B}_j^{(i)} \cup \mathcal{B}_{j'}^{(i')}) \tag{1}$$

両方のブックマーククラスタに含まれている Web ページの集合 $Y(\mathcal{B}_j^{(i)}, \mathcal{B}_{j'}^{(i')})$ を

$$Y(\mathcal{B}_j^{(i)}, \mathcal{B}_{j'}^{(i')}) = \mathcal{B}_j^{(i)} \cap \mathcal{B}_{j'}^{(i')} \tag{2}$$

と定義すると，比較するブックマーククラスタ同士の関係は図 2 のようになる．

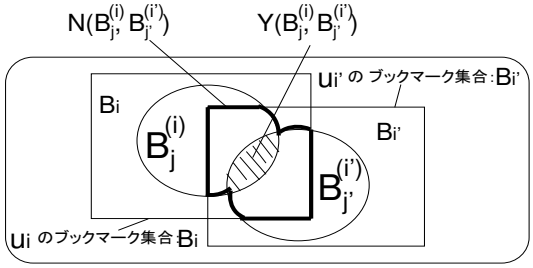

図 2. ブックマーククラスタの関係

$|N(\mathcal{B}_j^{(i)}, \mathcal{B}_{j'}^{(i')})| = n$, $|Y(\mathcal{B}_j^{(i)}, \mathcal{B}_{j'}^{(i')})| = y$ とする．このとき, 集合 $N(\mathcal{B}_j^{(i)}, \mathcal{B}_{j'}^{(i')})$ からサンプリングした Web ページが集合 $Y(\mathcal{B}_j^{(i)}, \mathcal{B}_{j'}^{(i')})$ に帰属する確率を l とみなすと，このサンプリングされる確率は二項分布 $L(n, y, l)$ として表現できる．このもとで, $\mathcal{B}_j^{(i)}$ と $\mathcal{B}_{j'}^{(i')}$ に類似性があるとみなせるときの l を l_1，$\mathcal{B}_j^{(i)}$ と $\mathcal{B}_{j'}^{(i')}$ に類似性がないとみなせるときの l を l_0 とし, $l_0 < l_1$ であるとする．この l_1,l_0 を用いて 尤度 $L(n, y, l_1)$ と $L(n, y, l_0)$ を比較し，どちらが大きいかを (3) 式から対数尤度比により, 判定する．

$$\log \frac{L(n, y, l_1)}{L(n, y, l_0)} = y \log \frac{l_1}{l_0} + (n - y) \log \frac{1 - l_1}{1 - l_0} \tag{3}$$

もし，l_0 という確率のもとでデータが観測されていれば, (3) 式は負の値になる．また, l_1 という確率のもとでデータが観測されていれば，(3) 式は正の値になる．この対数尤度比の値を利用して，ブックマーククラスタ同士の類似度 $\mathrm{sim}(\mathcal{B}_j^{(i)}, \mathcal{B}_{j'}^{(i')})$ は (4) 式のように算出される．

$$\mathrm{sim}(\mathcal{B}_j^{(i)}, \mathcal{B}_{j'}^{(i')}) = \max \left\{ \log \frac{L(n, y, l_1)}{L(n, y, l_0)}, 0 \right\} \tag{4}$$

類似度 $\mathrm{sim}(\mathcal{B}_j^{(i)}, \mathcal{B}_{j'}^{(i')})$ は対数尤度比が正の値をとれば, ブックマーククラスタ同士は類似しているとみなすため, その値を類似度とし, 対数尤度比が負の値をとれば, ブックマーククラスタ同士は類似していないとし, 類似度は 0 となる．

3.2.2 Web ページの推薦度算出方法

ユーザ u_i が, タグ $t_q^{(i)}$ に関する内容の Web ページの推薦を希望する場合, $t_q^{(i)} \in \mathcal{T}_i$ をクエリとして, 推薦システムに入力する．このタグ $t_q^{(i)}$ をクエリタグと呼ぶ．このとき, u_i の $t_q^{(i)}$ に対する Web ページ $p_w \in \mathcal{P}(w = 1, 2, \cdots, W)$ の推薦度を $R(t_q^{(i)}, p_w)$ とし, $R(t_q^{(i)}, p_w)$ が高い順に p_w を u_i に推薦する．$R(t_q^{(i)}, p_w)$ は (5) 式のように算出される．

$$R(t_q^{(i)}, p_w) = \sum_{i'=1}^{I} \sum_{j'=1}^{J_{i'}} a_{j'}^{(i')} \mathrm{sim}(\mathcal{B}_q^{(i)}, \mathcal{B}_{j'}^{(i')}) \tag{5}$$

$$a_{j'}^{(i')} = \begin{cases} 1 & (p_w \in \mathcal{B}_{j'}^{(i')}) \\ 0 & (p_w \notin \mathcal{B}_{j'}^{(i')}) \end{cases} \tag{6}$$

$R(t_q^{(i)}, p_w)$ は，$t_q^{(i)}$ に関するブックマーククラスタ $\mathcal{B}_q^{(i)}$ と p_w が属する全てのブックマーククラスタ $\mathcal{B}_{j'}^{(i')}(i' \neq i)$ との類似度の和として算出される．類似度の高いブックマーククラスタに多く含まれる Web ページは，ユーザの興味に合った Web ページと考えられ，推薦度は高くなる．

3.2.3 Web ページの推薦手順

以下の手順で Web ページを推薦する．推薦を受けるユーザ u_i のクエリタグ $t_q^{(i)}$ に関するブックマーククラスタをクエリブックマーククラスタ $\mathcal{B}_q^{(i)}$ とする．

Step1: u_i は $t_q^{(i)}$ を推薦システムに入力する．

Step2: (4) 式より，ユーザ u_i のクエリブックマーククラスタ $\mathcal{B}_q^{(i)}$ と他ユーザ $u_{i'}(i' \neq i)$ のブックマーククラスタ $\mathcal{B}_{j'}^{(i')}$ との類似度 $\mathrm{sim}(\mathcal{B}_q^{(i)}, \mathcal{B}_{j'}^{(i')})$ を算出する．

Step3: (5) 式より，Web ページ p_w の推薦度 $R(t_q^{(i)}, p_w)$ を $\mathcal{B}_q^{(i)}$ と p_w を含む全てのブックマーククラスタ $\mathcal{B}_{j'}^{(i')}$ との類似度の和として算出する．

Step4: $R(t_q^{(i)}, p_w)$ の高い上位 N 個の p_w を u_i に推薦する．

4 提案手法

4.1 着眼点

佐々木ら [4] の研究では，あるユーザ u_i がタグ $t_j^{(i)}$ を付与している u_i のブックマーク集合 $\mathcal{B}_j^{(i)} \in \mathcal{P}$ を，u_i の $t_j^{(i)}$ に関するブックマーククラスタとしており, $\mathcal{B}_j^{(i)}$ 内に含まれている Web ページの内容はスポーツ，映画など単一のトピックに関するものであることを前提としている．つまりブックマーククラスタは, ユーザ u_i のある単一のトピックに対する興味を表現している．この u_i が興味を示すトピックは，タグ $t_j^{(i)}$ で端的に表現されているといえる． 一般的に，複数のトピック全てで興味が類似するユーザは稀であり，あるトピックのみ興味が類似するユーザが大半である．例として, スポーツ，音楽，映画など複数のトピック全てで興味が類似するユーザは少ないが, スポーツというトピックのみで興味が類似するユーザは多い．ここで，ユーザが興味を示すトピックはさらに細分化され，複数のサブトピックに分かれてゆくという，トピックには個々のユーザ毎に異なる階層的な構造があると仮定する．例として，スポーツというトピックであれば，あるユーザは，野球やサッカー，別のユーザであれば，ラグビーや水泳など個々のユーザ毎に興味が異なるサブトピックに分割できると考えられる．しかし，佐々木らは

ユーザのサブトピックに対する興味を考慮しておらず，ブックマーククラスタ同士の類似度を算出する際，特定のサブトピックに関して部分的に類似するブックマーククラスタを考慮することで，さらに推薦精度が向上することが期待できる．

予備調査として, はてなブックマーク [1] からタグを 1 個以上付与したことがあるユーザ 6000 人をランダムに取得した結果，その中でタグを 2 個以上付与するユーザが全体の 60 %を占めた．つまり，ユーザは自分のブックマークに，複数のタグを付与する傾向があると考えられる．

この傾向から, ブックマーククラスタ $\mathcal{B}_j^{(i)}$ には，タグ $t_j^{(i)}$ と $t_j^{(i)}$ 以外の別のタグが付与されているブックマーク集合が部分集合として複数存在しているといえる．このような部分集合をサブブックマーククラスタと定義する．サブブックマーククラスタは, ユーザのサブトピックに対する興味を表現している．また, サブトピックは, u_i がタグ $t_j^{(i)}$ と良く組み合わせて使用するタグで表現されている．例として，図 3 のように, あるユーザの「スポーツ」というタグに関するブックマーククラスタ中に存在するサブブックマーククラスタを考える．ユーザは「スポーツ」というタグ以外にも「野球」,「サッカー」など「スポーツ」と関連するようなタグをそのブックマーククラスタ内の Web ページに付与しており，このようなタグが「スポーツ」というトピックの中の「野球」,「サッカー」というサブトピックを表現している．そして,「スポーツ」と「野球」が付与されているサブブックマーククラスタは,「スポーツ」の中の「野球」というサブトピックに対してユーザが興味を示しているということを表現している．

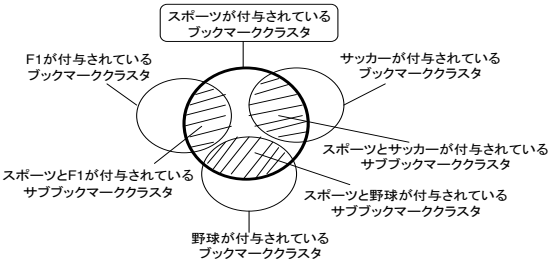

図 3. ブックマーククラスタとサブブックマーククラスタ

本研究では，ユーザのサブトピックに対する興味を考慮することで，特定のサブトピックに関して部分的に類似している興味を抽出し，推薦精度を向上させる手法を提案する．具体的には，サブトピックに対する興味を，ブックマーククラスタの部分集合であるサブブックマーククラスタで表現し，サブブックマーククラスタ同士の類似度を算出する．そして，類似しているサブブックマーククラスタに含まれる Web ページを推薦する．以下で，同一ユーザが利用するタグ同士の関連度の算出方法, サブブックマーククラスタ同士の類似度算出方法, Web ページの推薦度の算出方法を述べる．

4.2 タグ同士の関連度算出方法

ユーザ u_i が使用しているタグ $t_j^{(i)}$, $t_{j'}^{(i)}(j \neq j')$ の関連度 $rel(t_j^{(i)}|t_{j'}^{(i)})$ として以下の算出式を提案する．ただし, u_i が $t_j^{(i)}$ と $t_{j'}^{(i)}$ の両方を付与している u_i のブックマーク集合を $\mathcal{B}_{jj'}^{(i)}$ と定義する．

$$rel(t_j^{(i)}|t_{j'}^{(i)}) = \log TF(t_j^{(i)}, t_{j'}^{(i)}) \times ITF(t_j^{(i)}|t_{j'}^{(i)}) \tag{7}$$

$$TF(t_j^{(i)}, t_{j'}^{(i)}) = |\mathcal{B}_{jj'}^{(i)}| \tag{8}$$

$$ITF(t_j^{(i)}|t_{j'}^{(i)}) = \log \frac{|\mathcal{B}_i|}{|\mathcal{B}_{j'}^{(i)}| - |\mathcal{B}_{jj'}^{(i)}|} \tag{9}$$

$TF(t_j^{(i)}, t_{j'}^{(i)})$ が高ければ高いほど，u_i の中で $t_{j'}^{(i)}$ は $t_j^{(i)}$ と利用されることが多いタグといえる．$ITF(t_j^{(i)}|t_{j'}^{(i)})$ は値が高いほど，u_i の中で $t_{j'}^{(i)}$ が，複数のタグの中で特に $t_j^{(i)}$ と組み合わせて良く利用されるタグであることを示している．これらから $TF(t_j^{(i)}, t_{j'}^{(i)})$ と $ITF(t_j^{(i)}|t_{j'}^{(i)})$ が高いタグ $t_{j'}^{(i)}$ は，ユーザ u_i の中で $t_j^{(i)}$ と結びつきが強いタグであるといえる．

4.3 サブトピックタグの定義

ユーザ u_i のタグ $t_j^{(i)}$ に関するブックマーククラスタ $\mathcal{B}_j^{(i)}$ のサブトピックを表現するタグを，サブトピックタグとする．サブトピックタグとは，ユーザ u_i が $t_j^{(i)}$ と良く組み合わせて使用するタグであると考えられるため, 全ユーザの全ブックマークから $t_j^{(i)}$ との関連度が高いタグを $\mathcal{B}_j^{(i)}$ のサブトピックタグとして抽出する．ここで，u_i が使用しているタグを $t_{j'}^{(i)}(j' \neq j)$ としたとき，$t_j^{(i)}$ と $t_{j'}^{(i)}$ の関連度 $rel(t_j^{(i)}|t_{j'}^{(i)})$ が高い $t_{j'}^{(i)}$ がサブトピックタグである．

$t_j^{(i)}$ と抽出したサブトピックタグ $t_k^{(i)}$ が付与されている u_i のブックマーク集合を $\mathcal{B}_{jk}^{(i)}$ としたとき, $\mathcal{B}_{jk}^{(i)}$ が $\mathcal{B}_j^{(i)}$ のサブブックマーククラスタになる．

4.4 サブブックマーククラスタ同士の類似度算出方法

推薦を受けるユーザ u_i の $t_j^{(i)}$ に関するブックマーククラスタを $\mathcal{B}_j^{(i)}$ とすると, $rel(t_j^{(i)}|t_{j'}^{(i)})$ の値が大きい上位 Z 個の $t_{j'}^{(i)}$ をサブトピックタグ $t_k^{(i)}(k = 1, 2, \cdots, Z)$ とする．このときの $\mathcal{B}_j^{(i)}$ のサブブックマーククラスタを $\mathcal{B}_{jk}^{(i)}(k = 1, 2, \cdots, Z)$, あるユーザ $u_{i'}$ の $t_{j'}^{(i')}$ に関するブックマーククラスタ $\mathcal{B}_{j'}^{(i')}$ のサブブックマーククラスタを $\mathcal{B}_{j'k'}^{(i')}$ とする．ここで，ブックマーククラスタ $\mathcal{B}_j^{(i)}$ とサブブックマーククラスタ $\mathcal{B}_{j'k'}^{(i')}$ の類似度 $\mathrm{sim}(\mathcal{B}_j^{(i)}, \mathcal{B}_{j'k'}^{(i')})$ を以下の式で算出する．

$$\mathrm{sim}(\mathcal{B}_j^{(i)}, \mathcal{B}_{j'k'}^{(i')}) = \max_{1 \leq k \leq Z} \{\mathrm{sim}(\mathcal{B}_{jk}^{(i)}, \mathcal{B}_{j'k'}^{(i')})\} \tag{10}$$

一番興味が類似しているサブトピックを見つけるため, $\mathcal{B}_{j'k'}^{(i')}$ と $\mathcal{B}_{jk}^{(i)}(k = 1, 2, \cdots, Z)$ の類似度の中で最大類似度を $\mathrm{sim}(\mathcal{B}_j^{(i)}, \mathcal{B}_{j'k'}^{(i')})$ と定義する．なお，$\mathrm{sim}(\mathcal{B}_{jk}^{(i)}, \mathcal{B}_{j'k'}^{(i')})$ は，佐々木らの手法と同様に対数尤度比の考え方を用いており，サブブックマーククラスタ間で共通する Web ページの数を基に算出している．

4.5 Web ページの推薦度の算出方法

推薦を受けるユーザ u_i のクエリタグを $t_q^{(i)}$ とする．このとき, $t_q^{(i)}$ に対する Web ページ $p_w(w = 1, 2, \cdots, W)$ の推薦度 $R(t_q^{(i)}, p_w)$ が高い p_w を u_i に推薦する．$R(t_q^{(i)}, p_w)$ は，u_i のクエリブックマーククラスタ $\mathcal{B}_q^{(i)}$ と p_w が属する全てのサブブックマーククラスタ $\mathcal{B}_{j'k'}^{(i')}(i' \neq i)$ との類似度の和として，(11) 式のように算出される．

$$R(t_q^{(i)}, p_w) = \sum_{i'=1}^{I} \sum_{j'=1}^{J_{i'}} \sum_{k'=1}^{Z} a_{j'k'}^{(i')} \mathrm{sim}(\mathcal{B}_q^{(i)}, \mathcal{B}_{j'k'}^{(i')}) \tag{11}$$

$$a_{j'k'}^{(i')} = \begin{cases} 1 & (p_w \in \mathcal{B}_{j'k'}^{(i')}) \\ 0 & (p_w \notin \mathcal{B}_{j'k'}^{(i')}) \end{cases} \tag{12}$$

クエリブックマーククラスタと類似度の高いサブブックマーククラスタに多く含まれている Web ページは，ユーザの興味に合った Web ページと考えられ，推薦度は高くなる．

4.6 Web ページの推薦手順

提案手法では以下の Step で Web ページを推薦する．

Step1: 推薦を受けるユーザ u_i はクエリタグ $t_q^{(i)}$ を推薦システムに入力する．

Step2: (10) 式より，ユーザ u_i のクエリブックマーククラスタ $\mathcal{B}_q^{(i)}$ と他ユーザ $u_{i'}(i' \neq i)$ のサブブックマーククラスタ $\mathcal{B}_{j'k'}^{(i')}$ との類似度 $\mathrm{sim}(\mathcal{B}_q^{(i)}, \mathcal{B}_{j'k'}^{(i')})$ を算出する．

Step3: (11) 式より，p_w の推薦度 $R(t_q^{(i)}, p_w)$ を $\mathcal{B}_q^{(i)}$ と p_w を含む全てのサブブックマーククラスタ $\mathcal{B}_{j'k'}^{(i')}$ との類似度の和として算出する．

Step4: $R(t_q^{(i)}, p_w)$ の高い上位 N 個の p_w を u_i に推薦する．

5 評価実験

5.1 実験目的・方法

本評価実験では，本研究が提案する手法の有効性を示すため，佐々木らの従来手法 [4] と提案手法のブックマークの推薦精度の比較を行う．実験には，はてなブックマーク [1] のデータを使用した．ユーザ数を 3000 人，Web ページ数は約 140 万個となる．実験で想定するクエリブックマーククラスタは, 含む Web ページの数が多い上位 20 件のブックマーククラスタとする．この 20 個の各クエリブックマーククラスタ内に含まれる Web ページに共通に付与されているタグをそれぞれのクエリタグとし，クエリタグに対し, 推薦された Web ページの推薦精度を比較する．ユーザのクエリタグに対して，興味を満たした Web ページが推薦されているか否かを判断するための評価指標として推薦精度の定義は以下に示すとおりである．

$$\text{推薦精度} = \frac{\text{推薦された Web ページと正解データの被覆数}}{\text{推薦された Web ページ数}}$$

各々のクエリタグに対して推薦精度が算出されるため，最終的に 20 個の推薦精度が算出される．これらの 20 個の推薦精度の平均をとったものを平均推薦精度とし，従来手法と比較することで提案手法の有効性を示す．なお, 正解データとは, 各クエリブックマーククラスタ内に含まれる Web ページの中で, ブックマークした日付が新しい上位 150 件の Web ページとし, 正解データ以外の Web ページを学習データとしている．ブックマーククラスタ同士の類似度を算出する際には, 学習データに含まれる Web ページのみを利用し, その結果推薦される Web ページと正解データとの被覆数から推薦精度を算出する．なお, 各ブックマーククラスタに関して, 抽出するサブトピックタグは 20 個とする．

5.2 実験結果

図 4 は，推薦件数 N の値を 30，50,100 としたときの従来および提案に関しての平均推薦精度である．提案手法は全ての N に対して，従来手法と比べて高い推薦精度を達成している．このことから，提案手法の有効性を示すことができた．

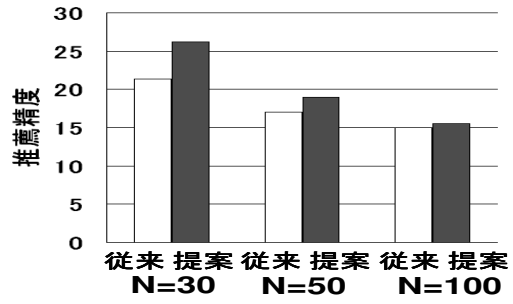

図 4. 各手法による推薦精度

5.3 考察

以下では例とし, クエリタグを「スポーツ」とした場合を述べる．表 1 は提案手法で抽出できたクエリタグのサブトピックタグである．値はクエリタグとの関連度を示している．

表 1. 「スポーツ」というクエリタグのサブトピックタグ

sports	1.753933	中国	0.177773
baseball	1.716548	訃報	0.161113
soccer	1.636098	すばらしい	0.158844
野球	0.407008	ほほえましい	0.135758
f1	0.352194	いい話	0.130489
mlb	0.27743	review	0.128723
サッカー	0.236898	cosplay	0.099991
football	0.225548	korea	0.091426
かっこいいぜ	0.22417	literacy	0.077948
event	0.208072	car	0.064575

抽出したサブトピックタグの上位には野球，サッカー，F1 などスポーツの中でもそのユーザが特に興味が強いサブトピックを表現しているようなタグが抽出できている．実際に，スポーツというクエリタグに対して，従来手法では，アイスホッケー，ラグビー，相撲などスポーツの中でも，幅広いサブトピックに関する内容の Web ページが網羅的に推薦されているのに対し，提案手法では野球，サッカー，F1 に関する内容の Web ページが集中的に推薦されていた．このことから，スポーツの中でも, ユーザが特に興味が強いサブトピックをうまく特定できたかつ, そのサブトピックに対して集中的に推薦したことにより提案手法の推薦精度が向上したと考えられる．

一方，全体的には推薦精度は向上したが, 中には従来と比較して推薦精度が低下するクエリタグも存在した．クエリブックマーククラスタ内に含まれる Web ページの数が少ないと, サブブックマーククラスタ内に含まれる Web ページの数が極端に少なくなってしまう．そのため，サブブックマーククラスタ同士の類似度を測る際に, 共通する Web ページが見つからなく, 類似しているサブブックマーククラスタを見つけられないため，精度が低下した．このことから, 提案手法では, クエリブックマーククラスタ内に含まれるブックマーク数がある程度大きい場合, つまり SBM を日々利用しているユーザに対しては有効であるといえるが, SBM の利用期間が少ないユーザには有効ではないといえる．

6 まとめと今後の課題

本研究では，ユーザの興味に応じた Web ページを推薦するために，SBM を利用した Web ページ推薦手法に着目し，ブックマーククラスタに含まれる Web ページの内容のサブトピックをタグで表現することで，特定のサブトピックに対する興味を抽出できる推薦手法を提案した．その結果，従来より高い推薦精度を示すことができた．今後の課題として，ブックマークに対してユーザが付けたコメントなどのタグ以外のメタデータの解析などが挙げられる．

参考文献

[1] 株式会社はてな. はてなブックマーク. http://b.hatena.ne.qp/

[2] del.ici.ous. http://www.delicious.com/

[3] 丹羽智史, 土肥拓生, 本位田真一, “Folksonomy マイニングに基づく Web ページ推薦システム,” 情報処理学会論文誌, Vol.47, pp.1382-1392, 2006.

[4] 佐々木祥, 宮田高道, 稲積泰宏, 小林亜樹, 酒井善則, “Soclal Bookmark におけるコンテンツクラスタ間の類似度を用いた web コンテンツ推薦システム,” 情報処理学会論文誌, Vol.48, pp.14-27, 2007.