個人ごとのタグ空間の類似性に基づく協調フィルタリング手法

白 川 貴 久*・倉 橋 節 也**

Personal Classification Space-based Collaborative Filtering in Social Bookmark Services

Takahisa SHIRAKAWA* and Setsuya KURAHASHI**

This paper improves Collaborative Filtering Algorithms (CF) to recommend similar bookmarks for users of Social Bookmark Services. Standard CF algorithms have a problem with precision in multi-genre cases such as social bookmarks. We aim at solving this issue and propose a personal classification space-based collaborative filtering algorithm. Our interpretation is that users' bookmarks are disposed in their own classification space made by tags. The bookmark is transformed into a scalar as a similarity to a starting bookmark for a recommendation. We compare the personal classification space with others' space using the similarities. We handle the similarities as the rating to the bookmark in standard CF algorithms. The proposed algorithm is compared with other existing algorithms. In conclusion, our algorithm shows better precision than others.

Key Words: Social Bookmark, recommendation system, Collaborative Filtering, multi-genre problem

1. はじめに

情報過多の解決は従来から研究されているが、インターネット上での情報発信の容易化とその情報の有用性の高まりに伴い、解決策の進化が求められている。

その解決策の1つとして、情報推薦システムの研究が盛んである。その主な手法として、ユーザ間の嗜好の類似性を利用する協調フィルタリング (Collaborative filtering, CF) がある。CF とは、推薦を受けるユーザの嗜好に類似した、ほかのユーザが好むアイテムを推薦する手法である。この CF は、Amazon の購買推薦など多くの実用例がある。その特長は、推薦の個人化と、コンテンツ非依存性にある。推薦の個人化とは、売上順位のように全利用者へ同じ推薦をするのではなく、過去の嗜好評価履歴などから各利用者個人に適合する推薦をすることである。そして、コンテンツ非依存性とは推薦コンテンツの中身に関する解析を必要としないことを指す。この2つの特長は、CF がユーザの嗜好の類似性を利用していることから生まれている。しかし、これまでの CF 手法では、多様なジャンルの情報が未整理で混在するような場合に、推

Fig. 1 Recommendation systems in social bookmark services

薦精度が低下する課題があった。著者はこれを multi-genre 問題と呼んでいる。

一方、ソーシャルブックマーク (Social BookMark, SBM) が、一般的に使われるようになってきている。これは、ブラウザのブックマーク管理を WWW 上で行なうことで、情報共有や協調的情報収集を行なうことができるサービスである。この SBM では、さまざまなジャンルのブックマークを容易に扱うため、ユーザがタグを付与する。たとえば「news」「面白い」「経営」などのキーワードがタグとして利用される。これにより、ユーザは多層的・非階層的な情報整理・分類を行なうことができる。また、SBM は URI (Uniform Resource Identifier) で表現できるインターネット上のさまざまなコンテンツを扱うことが本質的に可能である。すなわち、テキストで構成されたウェブページだけでなく、画像、音楽、動画、ウェブアプリなどが対象となる。この SBM 向けの推薦システムを考える場合、CF の特長が適する一方、multi-genre 問題の解決が必須である。

本研究は、CF を改良した SBM 向けの新しい情報推薦手法を提案する。この推薦手法では、SBM 上で登録された各ユー

* NEC パーソナルコンピュータ (株)
東京都品川区大崎 1-11-1

** 筑波大学大学院ビジネス科学研究科
東京都文京区大塚 3-29-1

* NEC Personal Computer, Ltd., 1-11-1 Osaki, Shinagawa-ku, Tokyo

** Graduate School of Business Sciences, University of Tsukuba, 3-29-1 Otsuka, Bunkyo-ku, Tokyo

(Received December 4, 2012)

(Revised May 7, 2013)

ザの嗜好情報の類似性に加えて情報整理方法の類似性にも基づいて推薦を行なう。それにより推薦の個人化を向上させることが目的である。想定するシステムは、Fig. 1 のように、SBM の利用者が BM を登録や閲覧したタイミングで、その関連情報が推薦される。

2. 関連研究

まず本研究の基礎をなす CF について概説する。CF の説明のためにつぎのような用語を用いる。アクティブユーザとは推薦を受けるユーザを、サンプルユーザとはアクティブユーザ以外のユーザを、アイテムとは推薦対象物を指す。CF は、ユーザ間の嗜好の類似性を利用する。具体的には、まず、アクティブユーザの近傍ユーザを見つける。近傍ユーザとは、アクティブユーザと同じアイテムをいくつか評価しており、その個人的嗜好の類似度が一定以上のサンプルユーザを指す。つぎに、その近傍ユーザの評価済アイテムへの嗜好評価値から、アクティブユーザの嗜好評価値を推定する。そして最後に、推薦アイテムを決定する。

つぎに CF を改善する関連研究を紹介する。従来からの研究として、cold-start 問題や sparsity 問題の対策が研究されている^{1)~7)}。これらの問題はアイテムの追加やアイテム数が多いことが原因で評価ユーザ数が不足し、近傍ユーザが見つけれないことに相当する。最近では、SBM への適応を想定して、いち早く BM を行なう能力をもつユーザ集合を発見し推薦元にするすることで、これらの課題を解決しようとする研究^{8)~10)}もあり、今後の進展が期待される。

また最近になって、本稿で multi-genre 問題と呼んでいる推薦精度が劣化する課題が指摘されている^{11)~23)}。これはジャンルごとに異なる近傍ユーザ集合を上手く識別できていないことに起因する。

ここで、上記の近傍ユーザ集合の識別の影響に関連し、高橋ら²¹⁾の研究を紹介する。高橋らは、標準的な CF 手法が有効に機能する範囲に関する指針を指摘している。標準的な CF 手法は、類似性について仮定をおいており、その仮定の当てはまりの度合いが精度に影響する。想定している類似性の仮定とは、

- 推薦システムが観測可能な「嗜好」情報において、被推薦ユーザ（アイテム）との類似度が高いほかのユーザ（アイテム）は、被推薦対象アイテム（ユーザ）との類似度も高い、

である。そして、その仮定について多くの研究で評価に用いられている映画の評価データセット (Movie Lens Data set) について分析を行ない、アイテムに非依存にユーザ間の相関が明らかなのは全体の 12%程度しかないことを報告している。すなわち、multi-genre 問題の把握をしていないと、88%のノイズの影響を受けるということになる。

2.1 multi-genre 問題の先行研究

まず、multi-genre 問題の先行研究および本研究の位置づけを Table 1 に整理する。分類のポイントは、ジャンルの推定方法 (カラム方向) と、分類の個人差の考慮レベル (ロウ方

Table 1 Related work and the position of this paper

	Not use tags	Use tags	Independent on tags
Classification scheme			This study
Genre	Jin Yang ¹¹⁾	Karen ¹³⁾	Sasaki ¹⁷⁾
None	Li Yu ¹²⁾	Niwa, Reyn ^{14), 15)}	Sugiyama ¹⁹⁾

向) である。本稿では、ジャンルの推定方法により 3 タイプに分類して紹介する。

タグを用いない方法 ジャンルを推定する手がかりがない場合を想定した研究^{11), 12)}が、これに属す。これらは、アイテム間の共起性からアイテムクラスタを推定する。

タグの表記を用いる方法 タグの表記に基づきクラスタリングを行なう研究^{13)~16)}が、これに属す。これらは、ユーザを跨って同一名称のタグは同一クラスタとし、さらにクラスタを結合することで、表記ゆれを吸収する。ただし、ユーザ間で共通に利用されるタグ名称は少なく、逆に少数もしくは一人のユーザしか付与しないタグ名称の種類が非常に多いため、名称を用いたクラスタリングは本質的に難しい。加えて、これらの研究では同名称異分野やユーザごとの概念の広狭のズレに基づくクラスタ分割については考慮できていない。このような影響が大きく、タグ名称を使うことの問題が指摘されている^{17), 20), 21)}。

タグの表記に非依存の方法 タグの名称には着目せず、ユーザ単位でのクラスタリングにのみ利用する研究^{17)~19)}が、これに属す。杉山ら¹⁹⁾は、ユーザごとのアイテム間類似度を同一タグが付与されている割合で表わし、それを加算することでアイテム間の類似度を求める手法を提案している。この方法の特長は、タグの利用をユーザごとのアイテム間類似度の計算に限定することで、ユーザ間でのタグの表記ゆれの問題を回避していることである。ただし、アイテム間の類似度は全ユーザにとって同等という仮説に従っており、分類に個人差があることを想定していないところに限界がある。佐々木ら¹⁷⁾は、タグは個人ごとに独立な分類を形成していると考えた。そして、2つのタグに関連するアイテムの共起性のみからタグ間の類似度という形で複数のタグクラスタを緩く結合している。佐々木らは情報整理方法に一定の個性を許容しているという面も評価できる。ただし情報整理の個性の考慮が不十分である。まず、2つのタグクラスタ間の類似度のみに着目し、ほかのタグの付与状態を含むユーザの全体的な情報整理の類似性は考慮していない。また、単独のタグでは必ずしも十分な類似度のタグが見つからない場合がある。その場合、複数のタグクラスタを組合せて集合演算を行ない、十分な類似度のタグクラスタを探索的に生成する必要がある¹⁸⁾。しかしながら、この処理は大変コストがかかり、効率的な生成方法の提案に至っていない。

以上のように情報推薦の活動の中で、ジャンルとはある 1 つの分野を指し、タグなしの場合はユーザプロフィール、タ

グ表記依存の場合は表記語の同義性、表記非依存の場合は共通に評価しているタグ間の類似性、などから算出したユーザの分類を意味している。たとえば、同じ趣味をもっているか、同じ名称のタグを使用しているか、共通の分類アイテムがあるか、などとなる。しかし、このジャンル分類では各ユーザの個性を正確に反映できない場合がある。人によりジャンルの捉え方、細かさは異なり、たとえば CF の論文を「論文」というジャンルと捉える人や、「ビジネスのヒント」と捉える人がいるかもしれない。また、従来の CF では、アイテムを複数のジャンルに分類していることを考慮していない。一方、本研究で提案する分類体系は、タグ表記とは関係なく、ユーザのアイテムへのタグ付けを個々人の分類空間への配置として捉えることで、個性をより正確に把握することを目指している。また、情報推薦において、状況に応じて推薦すべきアイテムが異なることは自然である¹⁸⁾。たとえば普段、工学書籍ばかりを購入している場合でも、いま着目している書籍が小説の場合には小説を探しているとの類推に基づき、その小説に類似した小説を推薦するという形で実装されることが想定できる。そこで本研究では、アクティブユーザが推薦直前に購入あるいは検索などによってアクセスした商品を、ユーザが「いま着目しているアイテム」と考え、これを起点アイテムと呼ぶ。この起点アイテムの導入は、推薦に対して個人の状況をもち込む効果が見込める。

2.2 記号定義

本節では、以降で用いる記号を定義する。

- ユーザ集合 $U \ni u$ 、アイテム集合 $I \ni i$ 、ユーザ u のタグ集合 $T_u \ni t$ 、アクティブユーザ u_a
- アイテム集合 I に対するユーザ u のタグ付与状況 $I \times T_u \rightarrow R \ni r(t, i)$
- 推薦起点アイテム i_0 へタグ t を付与しているサンプルユーザ集合 U_{sample}

$$U_{sample} = \{u \in U \mid \exists t \in T_u, r(t, i_0) = on\}$$

- 近傍ユーザ集合 $U_{neighbor} \in U_{sample}$ 、ユーザ u_j が評価しているアイテム I_j 、ユーザ u_j がアイテム i_k に付与したタグ集合 T_{jk}
- サンプルユーザ集合 U_{sample} がアイテム i_k に付与したタグ集合 $T_{sample \times k}$

$$T_{sample \times k} = \{t \in T_{u \in U_{sample}} \mid r(t, i_k) = on\}$$

と表わす。SBM では、ユーザ u の嗜好データ R はアイテム i について、多次元の 2 値データとなる。すなわち、複数のタグ T_u に関する、タグをつける/つけないという 2 値をとる。 $r(t, i) \in \{on, off\}$

2.3 佐々木の手法

本節では佐々木ら¹⁷⁾の手法を、本稿の提案手法に合わせて、推薦の起点となるアイテムのもつタグ集合から推薦する形に変形して紹介する。

佐々木らは、タグ間の類似度を確率モデルで定義している。

まず、両ユーザが共通に評価しているアイテムをサンプルデータと考え、2 つのタグに含まれるそれぞれのアイテム集合 (コンテンツクラス) への割り当て状況をベルヌーイ試行と考えモデル化する。そして、サンプルアイテムがいずれかのコンテンツクラスに含まれるサンプルアイテム数 n と、両方に含まれるサンプルアイテム数 k 、ユーザ a, b が評価しているアイテム I_a, I_b

$$n = |\{i \in I_a \cap I_b \mid r(t_a, i) = on \cup r(t_b, i) = on\}|$$

$$k = |\{i \in I_a \cap I_b \mid r(t_a, i) = on \cap r(t_b, i) = on\}| \quad (1)$$

を用いて、2 つのタグ t_a と t_b の類似度を、2 つのコンテンツクラスが一致するという尤度比検定として測る。すなわち、このコンテンツクラスの不一致仮説を p_0 、一致仮説を p_1 と設定すると

$$\begin{aligned} \text{sim}(t_a, t_b) &= \log \left(\frac{{}_nC_k \cdot p_1^k \cdot (1-p_1)^{n-k}}{{}_nC_k \cdot p_0^k \cdot (1-p_0)^{n-k}} \right) \\ &= k \cdot \log \frac{p_1}{p_0} + (n-k) \cdot \log \frac{1-p_1}{1-p_0} \quad (2) \end{aligned}$$

と表現される。なお、 p_0 と p_1 は、実験的に求めた値、 $p_0 = 0.1$ 、 $p_1 = 0.6$ を利用している。

そして、 $\text{sim}(t_a, t_b) > 0$ のサンプルユーザを近傍ユーザ $U_{neighbor}$ として利用する。ここで、アクティブユーザ u_a の未評価のアイテム i_c に対するタグ t_a の推定付与状況 $\overline{r(t_a, i_c)}$ 、および i_c の推定推薦度 $\overline{S(u_a, i_c)}$ は、

$$\overline{r(t_a, i_c)} = \sum_{t \in T_{neighbor \times c}} \text{sim}(t_a, t) \quad (3)$$

$$\overline{S(u_a, i_c)} = \sum_{t \in T_{a \times c}} \overline{r(t, i_c)} \quad (4)$$

となる。(3) 式は、サンプルユーザ集合 $U_{neighbor}$ がアイテム i_c に付与した全タグ $T_{neighbor \times c}$ と、アクティブユーザ u_a が i_c に付与したタグ t_a との類似度の単純和である。これは、 $T_{neighbor \times c}$ の和集合と t_a との類似度を計算していることに等しい。また、(4) 式は、アクティブユーザ u_a が i_c に付与した全タグに関する $\overline{r(t_a, i_c)}$ の単純和である。これは $T_{neighbor \times c}$ の和集合と $T_{a \times c}$ の和集合との類似度を計算していることに等しい。

3. 提案手法

3.1 分類空間ベース法

提案法では、1 つのタグをユーザ固有の分類空間を形成する 1 つの軸と考える。いいかえると、ユーザ u の分類空間は、ユーザ u が用いている全タグ $t_{uj}(j=0..n) \in T_u$ で張られている。そして、アイテム i へのタグづけを個々人の分類空間への配置と捉える。

$$\begin{aligned} \text{pos}(u, i) &= [r(t_{u1}, i), r(t_{u2}, i), \dots, \\ &\quad r(t_{uj}, i), \dots, r(t_{un-1}, i), r(t_{un}, i)] \quad (5) \end{aligned}$$

これにより、この分類空間への配置 $\text{pos}(u, i)$ はユーザの情

報整理方法の個性を陽に表現する。

異なるユーザの分類空間は、一般に、同じ軸はなく次元数も異なる。このため、アイテム i の特徴量がユーザ固有の分類空間上の位置ベクトル $pos(u, i)$ のままでは、異なるユーザ間で同じアイテム i に関する類似性を比較することができない。そこで、アイテム i の特徴量を、推薦起点アイテム i_o とのアイテム間類似度 $x(i, u, i_o)$ へ変換する。これは、スカラー値となる。

$$x(i, u, i_o) = Rel(pos(u, i), pos(u, i_o)) \quad (6)$$

ここで $Rel()$ は $\cos$ 距離などの変換関数を表わす。この変換により、異なるユーザ間で、推薦起点アイテム i_o とのアイテム間類似度 x という同じ軸での比較ができるようになる。そして、これは、各ユーザの分類空間について、推薦起点アイテム i_o に視点を置いた情報整理の類似性を比較しているといえる。

そして、ユーザ u の分類空間における、アイテム i に関する推薦起点アイテム i_o とのアイテム間類似度 $x(i, u, i_o)$ を、ユーザ u のアイテム i への評価値 $S(u, i)$ として、標準的なユーザベースの CF を行なう。ここで、Herlocker らの研究²⁾を参考に、ユーザ間の類似度の確からしさ ρ を導入する ((7) 式中では、 U_b に対する ρ として ρ_b と記載)。この ρ は、ユーザ間の評価アイテムの共有個数の対数である。対数をとっているのは、ある程度共有個数が大きければ十分確からしいので、徐々に頭打ちさせるためである。

なお、本論文では類似度による足しは行なわずすべてのサンプルユーザを近傍ユーザとして利用している。Herlocker らの報告では類似度による足しをするほうが推薦性能がよく、処理スピードも向上する。しかし、近傍を定義する類似度の閾値パラメータを導入すると、評価実験での調整箇所を増やしてしまう。今回はこれを避けた。

$$\begin{aligned} \text{sim}(u_a, u_b) &= \rho_b \times \frac{\sum_{i \in I} x(i, u_a, i_o) \cdot x(i, u_b, i_o)}{\sqrt{\sum_{i \in I} x(i, u_a, i_o)^2 \cdot \sum_{i \in I} x(i, u_b, i_o)^2}} \quad (7) \end{aligned}$$

$$\rho_b = \log(|\{i \in I_a \cap I_b\}|) \quad (8)$$

$$\bar{S}(u_a, i_c) = \frac{\sum_{u \in U_{\text{sample}}} \text{sim}(u_a, u) \cdot x(i_c, u, i_o)}{\sum_{u \in U_{\text{sample}}} \text{sim}(u_a, u)} \quad (9)$$

このような処理により、情報整理の個性すなわちタグ付与方法が類似している他ユーザからの推薦を受けることができる。以上が提案手法の本質的な手順である。

アイテム間類似度の変換関数 $Rel()$ を検討するにあたり、本稿では $\cos$ 関数（および $\cos$ に類似した関数）を想定した。これは、ユーザ間やアイテム間の類似度計算に常用されている。この基本的な類似関数を用いることで、提案方式の本質的な優位性を検証する。アイテム間類似度 $Rel()$ に $\cos$ 関数を利用することは、ユーザの分類空間についてつぎの3つの仮定をおくことになる。

仮定 1 各タグは直交

分類空間を張る各タグ T_u が互いに独立。

仮定 2 タグの重要度が正規化済

分類空間を張る各タグ T_u が互いに同重要度。

仮定 3 タグ上の分布が相似

分類空間を張る各タグ T_u 間で、アイテム分布が類似。

変換関数と精度の関係を検証することで、よりよい変換関数を決定する。このために、つぎの4種類のアイテム間類似度の変換関数 $Rel()$ を検討する。以降の数式の簡単化のため、

$$\begin{aligned} M_u &= |\{t \in T_u | r(t, i_o) = on \cap r(t, i) = on\}| \\ m_u &= |\{t \in T_u | r(t, i_o) = off \cap r(t, i) = off\}| \\ N_u &= |\{t \in T_u | r(t, i_o) = on \cap r(t, i) = off\}| \\ n_u &= |\{t \in T_u | r(t, i_o) = off \cap r(t, i) = on\}| \quad (10) \end{aligned}$$

という表記を用いることにする。なお、演算子 $|\{\cdot\}|$ は、集合の要素数を表わす。

cos 距離 [0, 1] $r()$ の値域を $\{1, 0\}$ とした $\cos$ 距離

$$\begin{aligned} x(i, u, i_o) &= \cos(pos(u, i), pos(u, i_o)) \\ &= \frac{\sum_{t \in T_u} r(t, i) \cdot r(t, i_o)}{\sqrt{\sum_{t \in T_u} r(t, i)^2 \cdot \sum_{t \in T_u} r(t, i_o)^2}} \quad (11) \end{aligned}$$

$$= \frac{M_u}{\sqrt{(M_u + n_u) \cdot (M_u + N_u)}} \quad (12)$$

cos 距離 [-1, 1] (11) 式で、 $r()$ の値域が $\{-1, 1\}$ の $\cos$ 距離

$$x(i, u, i_o) = \frac{M_u + m_u - N_u - n_u}{M_u + m_u + N_u + n_u} \quad (13)$$

一致重み付き cos 距離 (13) 式と類似しているが、同タグありでの一致となしでの一致、および不一致に重みをつけた。なお、重みは予備実験で探索的に求めた $K_{11} = 1.0$, $K_{00} = 0.01$, $K_{10} = K_{01} = 2.0$ を用いた。

$$x(i, u, i_o) = \frac{K_{11}M_u + K_{00}m_u - K_{10}N_u - K_{01}n_u}{M_u + m_u + N_u + n_u} \quad (14)$$

分布重み付き cos 距離 仮定 3 不成立を想定した変換関数。(12) 式に類似しているが、各タグのアイテム配置の重心 CI_t からの距離を重み $E \ni e$ として利用している。

$$CI_t = \frac{1}{|I_u|} \cdot \sum_{i \in I_u} r(t, i) \quad (15)$$

$$e_t = \max(|r(t, i) - CI_t|, |r(t, i_o) - CI_t|)$$

$$\begin{aligned} x(i, u, i_o) &= \cos(E \cdot pos(u, i), E \cdot pos(u, i_o)) \\ &= \frac{\sum_{t \in T_u} e_t \cdot r(t, i) \cdot e_t \cdot r(t, i_o)}{\sqrt{\sum_{t \in T_u} (e_t \cdot r(t, i))^2 \cdot \sum_{t \in T_u} (e_t \cdot r(t, i_o))^2}} \quad (16) \end{aligned}$$

Algorithm 1 に、擬似コードの形式で、分類空間ベース法の処理手順を示す。

Algorithm 1 The personal classification space-based collaborative filtering algorithm

```

100 全アイテムに対する推薦度を 0 に初期化
110  $u_a$  のブックマークしている各アイテムについて,
     $u_a$  のタグ空間での推薦起点  $i_0$  との
    アイテム間類似度ベクトル  $d_0$  を計算
120 for 推薦起点  $i_0$  をブックマークしている全ユーザ  $u_i$  について
{
130   $u_i$  のタグ空間で, アイテム間類似度ベクトル  $d_i$  を計算
140   $d_0$  と  $d_i$  について, 両者共通アイテムを取り出し,
     $D_0$  と  $D_i$  とする.
150   $\rho = \log(\text{ベクトル } D_0 \text{ の次元数})$ 
160   $w = D_0$  と  $D_i$  の cos 距離, を計算
170   $d_i$  に含まれるアイテム  $i_k$  の推薦度に  $d_i \times \rho \times w$  を加算
180 }
```

110 行目・130 行目のアイテム間類似度ベクトル (d_0 と d_i) の各要素は, 各アイテム i と推薦起点 i_0 とのアイテム間類似度 $x(i, u, i_0)$ を表わしている. これに基づいて, ユーザ間類似度が 160 行目で計算される.

140 行目では d_0 と d_i との両方に共通して現れるアイテムだけを, それぞれ D_0 と D_i に格納している. ここで, D_0 と D_i は同じアイテムに対応するアイテム間類似度が同じ要素番号になるように整列したベクトルである. このとき, D_0 と D_i の次元 (要素数) は同数であり, 150 行目で, その次元数の対数を計算している. これは 160 行目で求めているユーザ間類似度の確からしさを表わす係数 ρ である.

170 行目では, 通常の CF とほぼ同様に, (7) 式による u_i のアイテム評価値と, u_0 と u_i のユーザ間類似度から, (9) 式によって u_0 に対するアイテム推薦度を推定している. ここで, u_i のアイテム評価値は, そのアイテムと i_0 とのアイテム間類似度であり, u_0 と u_i のユーザ間類似度はその信頼係数 ρ での補正を含んでいる.

3.2 具体例による説明

Fig. 2 に示す具体例を用いて, 前節で説明した提案法の動作の流れについて説明を行なう. なお, 本節ではアイテム間類似度の計算値は cos 距離 $[0, 1]$ の場合で示す. 図は, アクティブユーザ u_a がアイテム i_0 を指定して類似アイテムの推薦を受ける場面を想定している. アクティブユーザ u_a は, タグ t_{a1} と t_{a2} を用いて, アイテム i_0 から i_5 までを分類している. また, 近傍ユーザの候補 u_1 と u_2 がそれぞれタグ t_{11} および t_{21} と t_{22} によって同じアイテム群を分類している. なお, Fig. 2 では, 実際の推薦処理を行なう場合に存在するはずの「 u_a がブックマークしていないアイテム」は省略されている.

まず 110 行目と 130 行目で, 各ユーザのタグ空間の特徴を表わす, アイテム i_0 に対する cos 距離ベクトルを求める. たとえば u_a の場合, タグ空間は (t_{a1}, t_{a2}) となり, $i_0 = (1, 0)$ や $i_1 = (1, 0)$ と表現できる. これら 2 つのアイテムの cos 距離を求めると, 1 となる. 同様に $i_0, i_1, \dots, i_5$ について計算すると結果的に, u_a の cos 距離ベクトルが $[1, 1, 1, 0, 0, 0]$, u_1 の cos 距離ベクトルが $[1, 1, 1, 1, 1, \times]$, u_2 の cos 距離ベクトルが $[1, 1, \times, 0, 0, \times]$ となる ($\times$ 印は対応するアイテムをブッ

Fig. 2 An example of a personal classification space in social bookmarks

クマークしていないため計算不可であることを示している).

つぎに, 140 行目で共通アイテムを取り出す. u_a と u_1 の場合, D_0 と D_1 は $[1, 1, 1, 0, 0]$ と $[1, 1, 1, 1, 1]$ となり, 同様に D_0 と D_2 は $[1, 1, 0, 0]$ と $[1, 1, 0, 0]$ となる. つぎに, 150 行目と 160 行目で, ユーザ間の類似度と信頼度を計算する. たとえば, w_1 は, $\cos(D_0, D_1) = \cos([1, 1, 1, 0, 0], [1, 1, 1, 1, 1]) = 0.77$ となる. ほかに, $\rho_1 = \log(5)$, $w_2 = 1$, $\rho_2 = \log(4)$ となる. この結果は, 図が示す u_a と u_2 が情報整理の類似性が高いことを, 正しく反映している. 最終的に 170 行目で u_1 と u_2 から推定した cos 距離ベクトル d_0 , すなわち推薦度は $[1, 1, 1, 0.45, 0.45, \times]$ となる.

3.3 精度測定尺度の提案

推薦精度は, 一般的には, 予測集合に対する recall と precision や, 平均絶対誤差 (MAE) などが広く用いられる^{3)~5)}. しかし, 少なくとも今回の用途では, 予測集合全体が参照されることはなく, 早い順位の推薦数個しか参照されない. このため, 予測が解集合に属すか否かを定める誤差評価や, MAE のような推薦度の誤差より, 順位の誤差のほうが重要である. また, 早い順位ほど適合結果を重視し, 遅い順位の適合結果は軽視するのが自然である. そこで, 本研究では, 下記の評価尺度 $DCG_{i^{th}}$ (Recall-DCG), $DCG_{p^{th}}$ (Precision-DCG) を用いる. (DCG: Discounted Cumulative Gain)

$$DCG_{i^0} = DCG_{p^0} = 0$$

$$DCG_{i^{th}} = DCG_{i^{th-1}} + \frac{|i^{th} - P(i^{th})|}{\gamma \cdot \log(i^{th})}$$

$$DCG_{p^{th}} = DCG_{p^{th-1}} + \frac{|I(p^{th}) - p^{th}|}{\gamma \cdot \log(p^{th})}$$

i^{th} が真の順位, p^{th} が予測順位, $I(p^{th})$ が p^{th} のアイテムの真の順位, $P(i^{th})$ が i^{th} のアイテムの予測順位, γ は正解集合の要素数と推薦集合の要素数の大きいほうの値を表わす.

DCG は順位 i のスコア値 S_i を用いて, 順位 p までの DCG_p を $DCG_p = DCG_{p-1} + \frac{S_p}{\log(p)}$ と表わすことができる. 一般に, スコア値は目的により異なる値を用いることができるため, Recall-DCG は, 真の順位を順位 i とし, その予測順位との差をスコアとして DCG を計算している. その際, 推薦集

合もしくは正解集合の要素数が大きいほど順位の差は大きくなる傾向があるため、 γ で除して正規化する。正解集合のうち推薦された集合の割合を求める Recall を、DCG の考え方を使得評価しているため、Recall-DCG と呼ぶことにする。

一方、Precision-DCG は、推薦順位を順位 i とし、その真の順位との差をスコアとして DCG を計算している。推薦集合のうち正解である集合の割合を求める Precision を、DCG の考え方を使得評価しているため、Precision-DCG と呼ぶことにする。また、一般的な Recall (再現率) と Precision (適合率) の関係同様に、トレードオフの関係にある評価指標の有利不利を含めて判断するために、単一の DCG ではなく、これら 2 種類の DCG を用いて評価を行なう。なお、同順位については、可能性のある提示順位をすべて考慮して誤差を求めている。

4. 実験

4.1 実験のデータと方法

実験には、「はてなブックマーク」からサンプリングした約 5000 人、約 240 万 URL のデータセットを用いた。また、推薦の起点アイテム i_0 とアクティブユーザ u_a とは、 i_0 のタグ付与数 λ がばらつくように、無作為に 40 点選択した。アイテムのタグ付与数 λ とは、アイテムに付与されたタグの数を数えたものである。 λ が少ないことは近傍ユーザが少ないことに直結し、推薦が困難になる可能性が高い。このため λ が 10 以下の範囲は意図的に細かくサンプリングした。それ以外は、できるだけ等間隔になるように、 λ の全範囲のデータを利用した。実際の推薦システムではアクティブユーザがブックマークしていないアイテムを推薦する。しかし、それでは推薦結果を客観的に評価することができない。そこで、本論文の実験は、アクティブユーザ u_a のアイテム i への正解推薦順位列を求めるために、タグ付けに基づく分類空間配置 $pos(u, i)$ を使得、(6) 式により各アイテム i と推薦起点アイテム i_0 とのアイテム間類似度 $x(i, u, i_0)$ ベクトルを計算する。つぎに、推薦対象アイテム i_c の順位を推定する際に、 i_c を除くほかのアイテムと起点アイテム i_0 との類似度に基づいて、ほかのユーザとの類似度を (7) 式によって計算し、それを重みとしてほかのユーザの i_c と i_0 との類似度を (9) 式によって計算することで、順位を求める。(7)、(9) 式にあるように、類似度に応じてユーザの影響が加味されることになるため、アクティブユーザと同じアイテムをブックマークするユーザが多いほど、正解順位の推薦精度が高まることになる。しかし、これは推薦システムのもつ本来の性質であり、正解となるアクティブユーザの推薦対象アイテム i_c が正解推薦順位列計算から除外されていることから評価の客観性は確保されており、問題はない。

4.2 実験結果

実データのサンプリングデータを用いた実験 A の結果を、Fig. 3 と Fig. 4 に示す。図は、横軸に順位 (真の順位、または推薦順位) をとり、縦軸に推薦精度 (Recall-DCG、または

Fig. 3 Recall-DCG 1-15th

Fig. 4 Precision-DCG 1-15th

Precision-DCG) をとり、最小値・第一四分点・中央値・第三四分点・最大値を示す箱ひげ図である。なお、順位を 15 位まで示すため、7 位までと 8 位以降の 2 つにグラフは分割されている。箱ひげ図は、佐々木方式、分類空間ベース法から 4 種のアイテム間類似度変換関数を比較しやすいように、これらを横に並べて図示している。アイテム間類似度の変換関数は左から、cos 距離 $[0, 1]$ ((12) 式)、cos 距離 $[-1, 1]$ ((13) 式)、一致重み付き cos 距離 ((14) 式)、分布重み付き cos 距離 ((16) 式) である。

5. 考 察

高橋²¹⁾は、標準的な CF では、ユーザ間の類似度が高い場合、ジャンル（アイテム集合）の好みの類似度も高い、との仮定に問題提起をし、それにより推薦精度が低下する問題が生じているとして、ジャンルの再検討が必要であることを報告した。これは、個々人にとってタグは個人固有のジャンルに相当し、そのようにタグを考慮する必要があることを示唆している。Nakamoto²⁴⁾は、一般的な CF の欠点として、今現在興味をもっている理由を考慮しないことを挙げ、タグを用いることでコンテキストを考慮することが可能であると主張している。そして、通常のブックマークの共起に基づく CF と比較し、タグを用いることで Precision が改善することを報告している。一方、佐々木¹⁷⁾は、タグの表記に依存する場合、表記ゆれや同音異義語、個人間でのタグの意味に差異があることを指摘し、タグ表記に非依存な提案手法がタグ表記に依存する推薦方法よりも Precision が改善することを示した。また Jaccard 係数で類似度を算出する一般的な CF でユーザ単位からタグ単位に置き換えた手法との比較も行ない、提案手法が良好な結果を得たことを報告している。このように、タグを考慮しない通常の CF やタグの表記に基づく手法に対して、タグの表記に依存しない佐々木の手法のほうが性能がよいことが先行研究で明らかとなっていることから、佐々木の方法と比較を行なうことで本研究の優位性を示す。

5.1 評価

本節では、実験結果に基づいて、分類空間ベース法と佐々木方式との比較、および 4 つのアイテム間類似度関数の比較をし、その要因を考察する。Fig. 3 と Fig. 4 を参照すると、提案方式の中では、特に、分布重み付き cos 距離が非常に高い精度を示している。また、Precision-DCG については、cos 距離 $[-1, 1]$ もよい値を示している。

一致重み付き cos 距離については、提案方式の中では最も精度が悪い結果であった。ここで、一致重み付き cos 距離について補足しておく。一致重み付き cos 距離は、予備実験を行なって、試行錯誤的に、3.1 節に示した重み $K_{11} = 1.0$, $K_{00} = 0.01$, $K_{10} = K_{01} = 2.0$ を求めた。その際には、cos 距離 $[-1, 1]$ の精度を大きく改善し、重み付き cos 距離と並ぶ結果を得ていた。しかし、本実験では提案方式の中で、最も性能が悪い結果となった。これは、重みのパラメータチューニングがデータに非常に依存していることを表わしている。

つぎに、同じようにタグ表記に非依存な手法を用いる佐々木方式との比較を考察する。分類空間ベース法は全般的に佐々木方式より精度がよい。単純な cos 距離 $[0, 1]$ を用いた場合でも Precision-DCG は佐々木方式と同等、Recall-DCG は佐々木方式より小さく抑えられている。

提案方式のアイテム間類似度への変換は分類空間全体の情報を 1 次元空間へ縮退したものである。一方、尤度比検定を用いる佐々木方式では、2 つのタグクラス間の類似度のみに着目し、ほかのタグの付与状態を含むユーザの全体的な情

報整理の類似性は考慮していない。また、タグクラス間の類似性の検討の際、個々のアイテムのタグクラスへの帰属度合いの違いを考慮していない。これは、分類空間を空間としてではなく、軸単独で扱い、着目している軸ごとに類似する軸を探索していることに相当する。それに対し本研究では、ユーザがアイテムに付与した全タグによる分類空間を使って、そのユーザが認知しているすべてのアイテム間の距離を求めている。このように、本提案方法は、情報整理の個性が、個人差が大きいアイテム表記への解釈を使わず、各個人のタグ付けという情報から「アイテムを分類する個性 = 情報整理の個性」を反映していることになる。以上の理由から、佐々木方式に比べ、提案方式は情報整理の個性がより反映され、その効果が精度の向上をもたらしたと推定する。

また、佐々木方式に比べ、提案方式は利用するタグ軸が多い。この視点で、数式による説明を試みる。2.3 節に示した (1) 式と、3.1 節に示した (10) 式を比較参照する。これらの式は、互いに計測しているのがアイテムの個数と、タグの個数の違いがあるので、佐々木の n と k と、分類空間ベース法の M_u, m_u, N_u, n_u を相互に変換することができず、直接比較することはできない。しかし、計数の元になるユーザ id・アイテム id・タグ id の 3 項の組で考えると、情報量の大小関係は比較できる。ここで M_u, m_u, N_u, n_u は、それぞれ (10) 式におけるアイテム間のタグ付与状況の要素数である。佐々木方式は、ユーザ u_a の情報として M_a と N_a に相当する情報を、ユーザ u_b の情報として M_b と N_b に相当する情報を利用している。しかし、 m_a と n_a 、および m_b と n_b の情報は使っていない。これに対して、cos 距離 $[0, 1]$ は、 $M_a \cdot N_a \cdot M_b \cdot N_b$ に加え、 n_a と n_b の情報を使っている。精度の差はここから来ていると考える。

さらに、cos 距離 $[-1, 1]$ の場合、 $M_a \cdot m_a \cdot N_a \cdot n_a \cdot M_b \cdot m_b \cdot N_b \cdot n_b$ のすべての情報を使っている。このため、cos 距離 $[0, 1]$ より精度が向上している。ただし、cos 距離 $[0, 1]$ は、精度が悪化する場合があり、その原因を解析すると、 $M \cdot m \cdot N \cdot n$ のうち、 m の数が支配的で、ほかの情報の特徴が覆い隠されているためであった。これは、非常に多くのタグを使用し（最大 20,351 個/人）、同じタグは 1・2 回（中央値は 1 回）しか付与されていないことが多いためである。実際に、cos 距離 $[-1, 1]$ の値で計算した多くのアイテム間類似度が約 0.9999 となっていた。

一致重み付き cos 距離は、上記のような場合に、cos 距離 $[-1, 1]$ の場合に支配的であった m の重みを下げ、ほかの情報の特徴を浮き上がらせた効果を期待し設計した。しかし、上述のとおり、一致重み付き cos 距離は、先に述べたように重みパラメータがデータ依存性が高く、機能しなかった。

分布重み付き cos 距離は、佐々木方式と比較し、Precision-DCG と Recall-DCG の両面で明らかに精度がよい。これは各タグごとのアイテムの分布状況を求め、タグの付与の有無の珍しさを評価することができるためである。先ほどと同じ方法で説明すると、(15) 式の計算中で、すべてのアイテム

(佐々木方式で利用していない, m_a と n_a および m_b と n_b の条件に相当するアイテムも含む) の情報を活用して, 分布状況を求めている. これにより, 個々のアイテムにとってのタグの重要度を考慮することができ, 精度が向上したと推測する.

結論としては, 今回の実験データの範囲では, 佐々木方式に比べ分類空間ベース法は精度がよく, 特に分布重み付き cos 距離の場合には非常に優位であった.

5.2 提案手法の計算コスト

本節では, 分類空間ベース法の計算コストについて考察する. 処理時間の大半を占める探索起点アイテムとの類似度計算 ((6) 式) は事前に行なうことができる. その場合, 事前の計算ではアイテム間の全組合せについてアイテム間類似度を求めておくこととなる. それは, ユーザごとのブックマークの個数の階乗回の計算コストとなる. しかし, その処理はアイテムの登録時やタグの設定・変更時などに分散することができる. その場合, 1つのアイテムとほかのすべてのアイテムとのアイテム間類似度を求めるだけでよい. これにより, 一度に行なう計算は, 高々, 各ユーザのブックマークの個数回の計算コストとなる. このような処理手順にすることで, 推薦提示の応答速度を確保することができる.

これに対し, アイテムの登録時やタグの設定・変更時にタグ付け情報をもとに, 佐々木方式¹⁸⁾のアイテム集合を更新する場合, 非常に大きな計算コストが必要である. すなわち, 更新されたタグについて, ほかのユーザのタグを含む, すべてのタグとの間で集合演算を行ないアイテム集合の更新を行なう必要がある. また, 集合演算も和集合演算, 差集合演算, 補集合演算の考慮が必要である. さらに, この集合演算はタグ2つだけによるとは限らない. 3つの場合や4つの場合など, 非常に多くの組合せを想定する必要がある. これはユーザおよびタグの増加にともない莫大な計算コストを必要とする.

6. おわりに

本稿では, さまざまなジャンルの Web ページを対象にするソーシャルブックマークを用いた情報推薦システムにおいて, 標準的な CF 手法では精度のよい推薦ができない問題に取り組んだ. そして, これを解決する推薦手法として分類空間ベース法を提案した. 先行研究との比較により, 分類空間ベース法の優位性を確認し, アイテム間類似度関数に分布重み付き cos 距離を用いた場合には, ほぼすべての実験において先行研究に対し Recall-DCG 値, Precision-DCG 値の両面で大きな優位性を確認した. 最後に, 今後の課題を下記に挙示し, 本稿をむすぶ. 1) アイテム間類似度関数の設定時に前提とした, 分類空間の直交性, タグの重要性の正規性, およびタグごとのアイテムの配置分布の相似性に関し, 実データの特性を分析する必要がある. 2) 4種類のアイテム間類似度関数について, 上記の特性を乱した場合の影響について検証を行なう必要がある. 3) パラメータとデータとの関係性の解析と, チューニング手法についての検討が必要である. 理想的には, パラメータ自体もデータに自動適応するようなア

ルゴリズムに改良していくべきである. たとえば, 一致重み付き cos 距離のパラメータは, 予備実験における試行錯誤の結果で決定した. これに対し, 最良の結果であったアイテム間類似度関数の分布重み付き cos 距離関数は, アイテム分布の重心から各タグの重みを計算している. 後者のような改良が有用である.

参 考 文 献

- 1) J.S. Breese, D. Heckerman and D. Kadie: Empirical Analysis of Predictive Algorithms for Collaborative Filtering, Proceedings of the 14 Conference on Uncertainty in Reasoning, 43/52 (1998)
- 2) J.L. Herlocker, J.A. Konsta, A. Borchers and J. Riedl: An Algorithmic Framework for Performing Collaborative Filtering, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 230/237 (1999)
- 3) 神島敏弘: 推薦システムのアルゴリズム (1), 人工知能学会誌, **22-6**, 826/837 (2007)
- 4) 神島敏弘: 推薦システムのアルゴリズム (2), 人工知能学会誌, **23-1**, 89/103 (2008)
- 5) 神島敏弘: 推薦システムのアルゴリズム (3), 人工知能学会誌, **23-2**, 248/263 (2008)
- 6) X. Li and T. Murata: Exploiting Formal Concept Analysis in a Customizing Recommendation for New User and Gray Sheep Problems, 電気学会論文誌 C, 電子・情報・システム部門誌, **132-5**, 782/789 (2012)
- 7) Q. Liu, E. Chen, H. Xiong, C.H.Q. Ding and J. Chen: Enhancing Collaborative Filtering by User Interest Expansion via Personalized Ranking, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, **42-1**, 218/233 (2012)
- 8) 百田, 伊東: ソーシャルブックマークに基づく情報発見, Data Engineering Workshop, 11-15 (2008)
- 9) 大力, 大向, 武田: ソーシャルブックマークにおけるイノベータに注目した情報推薦手法の提案, The 22nd Annual Conference of the Japanese Society for Artificial Intelligence, 2D3-3 (2008)
- 10) 毛受, 吉川: ブックマークの時系列情報を利用したソーシャルブックマークにおける注目度予測, Data Engineering Workshop, B9-5 (2008)
- 11) J. Yang and K.F. Li: An Adaptive User-Genre-Item Model for Collaborative Filtering, The IEEE Pacific Rim 2007 Conference on Communications, Computers and Signal Processing, 257/262 (2007)
- 12) L. Yu: Using Ontology to Enhance Collaborative Recommendation Based on Community, Proceedings of 9th International Conference on Web-Age Information Management, 45/49 (2008)
- 13) H.L. Karen and K. Tso-Sutter: Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms, Proceedings of 23rd Annual ACM Symposium on Applied Computing, 16/20 (2008)
- 14) S. Niwa, T. Doi and S. Honiden: Web Page Recommender System based on Folksonomy Mining, Proceedings of the Third International Conference on Information Technology: New Generations, 388/393 (2006)
- 15) R. Nakamoto, S. Nakajima, J. Miyazaki and H. Kato: Live-Updating Website Recommendations Using Reasonable Tag-based Collaborative Filtering, Data Engineering Workshop, B9-4 (2008)
- 16) X. Shi, F. Chen, M. Toyoda, M. Wang and M. Kitsuregawa: Tag Recommendation in Photo Sharing Ser-

- vices based on Multi-Granular Context Dependency, Proceedings of the 4th Forum on Data Engineering and Information Management, B3-1 (2012)
- 17) 佐々木, 宮田, 稲積, 小林, 酒井: Social Bookmark におけるコンテンツクラス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データベース, **48**-SIG 20, 14/27 (2007)
- 18) 佐々木, 宮田, 稲積, 小林, 酒井: コンテンツクラスタの論理演算を導入したコンテンツ推薦, Data Engineering Workshop, A2-8 (2007)
- 19) 杉山, 関, 青野: ユーザのタグ付けの傾向を利用したソーシャルブックマーク内の関連ページ検索手法, Data Engineering Workshop, B9-2 (2008)
- 20) J. Voss: Tagging, Folksonomy & Co-Renaissance of Manual Index?, Proceedings of the 10th International Symposium for Information Science, Technical Report, abs/cs/0701072 (2007)
- 21) 高橋, 小林: MovieLens Data にみる協調フィルタリングの失敗, 情報処理学会研究報告データベースシステム, 2009-26 (2009-DBS-149), 1/8 (2009)
- 22) C. Sanden and J.Z. Zhang: Enhancing multi-label music genre classification through ensemble techniques, Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, 705/714 (2011)
- 23) D. Brezeale and D.J. Cook: Automatic Video Classification: A Survey of the Literature, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, **38**-3, 416/430 (2008)
- 24) R. Nakamoto, S. Nakajima, J. Miyazaki and S. Uemura: Tag-Based Contextual Collaborative Filtering, IAENG International Journal of Computer Science, **34**-2, IJCS.34.2.08 (2008)

=====

[著 者 紹 介]

白 川 貴 久

2010 年筑波大学大学院ビジネス科学研究科経営システム科学専攻修士 (経営システム科学). 現在 NEC パーソナルコンピュータ (株). 商品企画, 推薦システム開発などに従事.

倉 橋 節 也 (正会員)

2002 年筑波大学大学院経営・政策科学研究科企業科学専攻博士 (システムズ・マネジメント). 2006 年筑波大学大学院ビジネス科学研究科助教授, 現在ビジネスサイエンス系准教授. 2009 年 University of Groningen (オランダ) 客員研究員. 経営システム分析, 社会シミュレーションなどの研究に従事.

=====