

## SORNによるコンテンツ推薦システム

## Content Recommendation System using SORN

清政 貴文†

六井 淳†

Takahumi Seimasa

Jun Rokui

## 1. まえがき

Webの情報は日々蓄積され、利用できるコンテンツの数は増加している。一方、膨大なWebページの中から必要な情報や興味のあるコンテンツを探すことは困難である。ユーザの負担を減らしてWebの利便性を高めるため、コンテンツ推薦システムの重要性が高まっている[1]。

コンテンツ推薦システムはコンテンツベース型と協調フィルタリング型、両者のハイブリッド型の3通りに大別できる[1]。コンテンツベース型はコンテンツの特徴量を扱い、ユーザの嗜好モデルを構築して推薦を行う。協調フィルタリング型は他のユーザの行動履歴を参照し、類似ユーザの利用コンテンツを推薦する。ハイブリッド型は両手法を組み合わせたものである。

コンテンツベース型は推薦の質がサービスの利用者数に影響されず、利用者の居ない新規のコンテンツを推薦できる。課題として推薦内容の偏り、メディアの種類によっては特徴量抽出が困難なことが挙げられる[2]。

協調フィルタリング型はコンテンツの特徴量解析が必要無く、異なるメディア同士を結び付けることができる。課題として履歴の蓄積が少ないシステム稼働当初の推薦精度が落ち、新規コンテンツの推薦が難しいといったことが挙げられる[2]。

コンテンツベース型と協調フィルタリング型の問題は相補的な関係にある[3]。両者の利点を活かして推薦の質を高めようとするのがハイブリッド手法である。

本論文ではコンテンツベース型、協調フィルタリング型として運用でき、インタラクティブ性を取り入れてユーザ満足度向上を図る推薦システムを提案する。

## 2. 提案手法

## 2.1 SORN

Self-Organizing Relationship Network(SORN)という学習器を用いる。SORNは堀尾[4]らによって提案されたSelf-Organizing Map(SOM)[5]の拡張手法である。SORNは好ましい入出力関係を自己組織化するために用いられる。SORNの構成を図1に示す。入力層、出力層、競合層から成り、それぞれ $n$ 、 $m$ 、 $N$ 個のユニットをもつ。競合層上の $i$ 番目のユニットは入力層、出力層の全ユニットとそれぞれ結合重みベクトル $\mathbf{W}_i = (W_{i1}, \dots, W_{ij}, \dots, W_{im})$ と $\mathbf{U}_i = (U_{i1}, \dots, U_{ik}, \dots, U_{im})$ を介して結合している。 $i$ 番目のユニットの結合重みベクトルを $\mathbf{V}_i = (\mathbf{W}_i, \mathbf{U}_i)$ とする。結合重みは $[0, 1]$ の値をとる。

学習用ベクトルとして入出力ベクトルの組 $\mathbf{I} = (\mathbf{X}, \mathbf{Y})$ を与え、この組に対する評価値 $E$ も同時にネットワークに与え

る。入出力ベクトルの各値は $[0, 1]$ の値に正規化される。評価値は $[-1, 1]$ の値をとり、正の値は望ましい入出力関係を、負の値は悪い関係を表す。

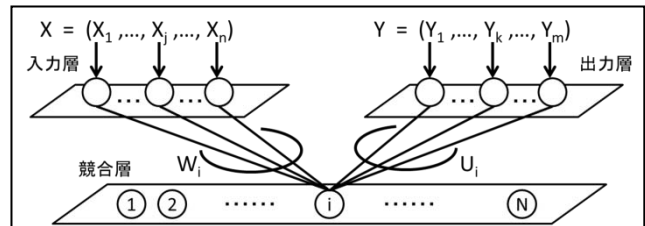


図1 SORNの構造

入出力ベクトル $(\mathbf{X}, \mathbf{Y})$ と評価値 $E$ が与えられたとき、式(1)(3)で入力層、(2)(4)で出力層との結合重みを更新する。 $W_{ij}$ 、 $U_{ik}$ は $i$ 番目の競合層ユニットがもつ $j$ 番目の入力層ユニット、 $k$ 番目の出力層ユニットとの結合重みである。 $\mathbf{I} = (\mathbf{X}, \mathbf{Y})$ 、 $\mathbf{X} = (X_1, \dots, X_j, \dots, X_n)$ 、 $\mathbf{Y} = (Y_1, \dots, Y_k, \dots, Y_m)$ の学習では、評価値 $E$ が正なら $W_{ij}$ 、 $U_{ik}$ は $X_j$ 、 $Y_k$ に近づき、負なら遠ざかる。学習データと競合層の差が大きければ大きく、小さければ小さく更新する。 $\alpha(t)$ は学習の収束用関数であり、学習の進行度 $t$ に応じて重みの更新値を小さくする。 $\beta$ は評価値 $E$ が負の場合の更新値を制御し、 $\beta$ が小さければ小さく、大きければ大きく遠ざける。 $\text{sgn}()$ は符号関数であり、正の値には1、負の値には-1を返す。

$$\begin{cases} \text{if } 0 \leq E \\ W_{ij}^{\text{new}} = W_{ij} + \alpha(t) \cdot E \cdot (X_j - W_{ij}) & (1) \\ U_{ik}^{\text{new}} = U_{ik} + \alpha(t) \cdot E \cdot (Y_k - U_{ik}) & (2) \\ \text{else if } 0 > E \\ W_{ij}^{\text{new}} = W_{ij} + \alpha(t) \cdot E \cdot \exp\left(-\frac{|E-V|}{\beta}\right) \cdot \text{sgn}(X_j - W_{ij}) & (3) \\ U_{ik}^{\text{new}} = U_{ik} + \alpha(t) \cdot E \cdot \exp\left(-\frac{|E-V|}{\beta}\right) \cdot \text{sgn}(Y_k - U_{ik}) & (4) \end{cases}$$

学習後のSORNは入力 $\mathbf{X}$ に対し式(5)を計算することで、 $\mathbf{X}$ に対する好ましい出力 $\mathbf{Y}$ を得ることができる。 $z_i$ は $i$ 番目の競合層ユニットと入力 $\mathbf{X}$ の類似度であり、 $k$ 番目の出力 $y_k$ は、 $U_{ik}$ の $z_i$ による重み付き平均和で求められる。式(5)で $\mathbf{Y}$ を求めることは、入力 $\mathbf{X}$ に対して $N$ 個のファジィif-thenルールによるファジィ推論を行うことを意味する[4]。式(6)中の $\gamma$ はファジネスを表し、 $\gamma$ が大きければ結合重みが出力に反映されやすくなる。

$$\begin{cases} y_k = \frac{\sum_{i=1}^N z_i U_{ik}}{\sum_{i=1}^N U_{ik}} & (5) \\ z_i = \exp\left(-\frac{|X - W_{i1}|}{\gamma}\right) & (6) \end{cases}$$

†(法) 島根大学 総合理工学研究科

提案手法では入力層にユーザ情報, 出力層にコンテンツ情報を与え, 獲得した結合重みを推薦に利用する.

## 2.2 コンテンツベース型

コンテンツの内容やタグから生成したコンテンツ特徴ベクトルを  $\mathbf{C}=(C_1, C_2, \dots, C_m)$  とする. ユーザのコンテンツ利用履歴, 及び年齢や性別などのデモグラフィック情報から生成したユーザの属性ベクトルを  $\mathbf{U}=(U_1, U_2, \dots, U_n)$  とする. コンテンツベース型ではこの両ベクトルの関係を SORN に学習させる(図2). 属性  $\mathbf{U}$  をもったユーザが特徴  $\mathbf{C}$  をもつコンテンツを利用して評価  $E$  を下した場合, 学習データ  $(\mathbf{U}, \mathbf{C}, E)$  が生成される. 学習データは特定期間毎に収集して更新し, 流行を反映させた学習を行う.  $E$  が正のデータの学習を行う場合,  $(\mathbf{U}, \mathbf{C})$  に最も近い結合重みをもつ競合層ユニットとその周辺ユニットの重みを  $(\mathbf{U}, \mathbf{C})$  に近づけ,  $E$  が負の場合は遠ざける. 学習を繰り返し, 各ユーザ層が何のコンテンツを評価したかの関係を競合層に獲得させる.



図2 コンテンツベース型の学習

学習後の競合層の  $i$  番目のユニットには  $C_i$  に近い特徴をもつコンテンツ群を対応させる. 同様に全てのユニットにコンテンツ群を対応させ, 図3のようなコンテンツマップを作成する. マップの各マスの色は保持するコンテンツ情報により異なる.  $i$  番目のユニットが対応したマスについて  $U_i$  とシステム利用ユーザ属性  $\mathbf{U}$  の類似度を求め, マップ上での表示を変化させる(図4). 類似度はユーザベクトル間の距離を基に算出する. ユーザはマップに対して操作することで, コンテンツの推薦を受けられる. マップの各マスを選択して対応したコンテンツ群を表示する. また, 特定のマスを選択して類似度を再計算することで, 選択したマスのユーザ属性を反映したマップを表示する. なお, コンテンツマップは学習対象とした全ユーザで共有する. ユーザ操作を伴わない推薦として, ユーザ属性  $\mathbf{U}$  の入力に対して式(5)に基づいて得られた  $\mathbf{C}$  に近い特徴をもつコンテンツを表示する.

この手法ではユーザ個人の嗜好モデルではなく, ユーザ全体の嗜好モデルを SORN 上に構築する. そのため, 学習後の競合層解析により学習データ収集期間中における全体傾向の抽出が期待できる. また, マップ操作により他者の興味を反映した推薦を受けることができ, 思いがけない発見的な推薦が可能になる.

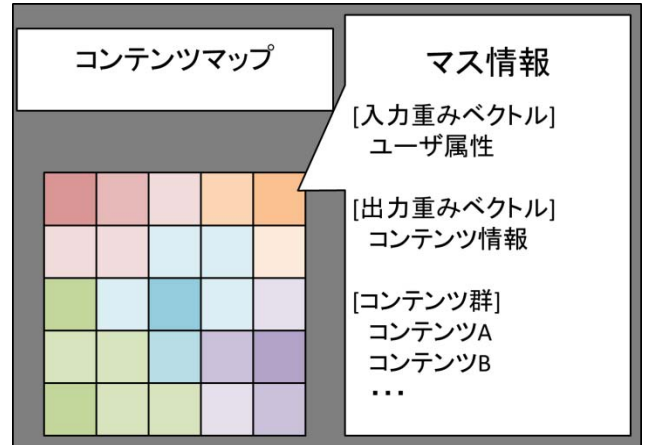


図3 コンテンツマップの構造



図4 マップへのユーザ情報反映

## 2.3 協調フィルタリング型

コンテンツベース型の構成を協調フィルタリング用に修正する. SORN の入力層にはコンテンツの利用履歴を直接与え, 出力層にも利用コンテンツをそのまま与える. 競合層ユニットは入力重み集合  $\mathbf{I}=\{[I_1, I_{w1}], [I_2, I_{w2}], \dots, [I_n, I_{wn}]\}$ , 出力重み集合  $\mathbf{O}=\{[O_1, O_{w1}], [O_2, O_{w2}], \dots, [O_m, O_{wm}]\}$  をもつ. 集合中の組の  $I_j, O_k$  はコンテンツの ID,  $I_{wj}, O_{wk}$  は各コンテンツに対する重みを表す. これは総コンテンツ数を  $N_c$  としたとき,  $N_c$  個の入力層・出力層ユニットの内, それぞれ  $n$  個・ $m$  個と結ばれている. ユーザのコンテンツ利用履歴を  $\mathbf{L}=(C_1, C_2, \dots, C_A)$  とする. 履歴  $\mathbf{L}$  をもったユーザがコンテンツ  $C_{USE}$  を利用して評価  $E$  を下した場合, 学習データ  $(\mathbf{L}, C_{USE}, E)$  が生成される. 学習時は類似度  $S$  が最も高かった競合層ユニットと近傍ユニットについて, 評価値  $E$  に加え入力集合として  $\{[C_1, 1], [C_2, 1], \dots, [C_A, 1]\}$ , 出力集合として  $\{[C_{USE}, 1]\}$  を与えて重みの更新を行う. ここでは  $C_{USE}$  に対する評価のみ考慮し, 利用履歴  $\mathbf{L}$  中のコンテンツに対する過去の評価は考えないものとする. 学習データと競合層ユニットの類似度  $S$  は式(7)で求める.

$$S = \sum_{I_j \in L \cap I_c} I_{wj} + O_{wc} \quad (7)$$

$O_{w_c}$ の値は  $O_j = C_{USE}$  となる組が  $\mathbf{O}$  に含まれている場合  $O_{w_j}$ , そうでなければ 0 となる. 競合層ユニットの入力重み集合  $\mathbf{I}$  の組に含まれないコンテンツ  $C_i$  について  $C_{w_i}=0$  として学習する. 同様に  $\mathbf{O}$  に含まれない  $C_{USE}$  について,  $[C_{USE}, 0]$  が  $\mathbf{O}$  に含まれているとする(図 5).

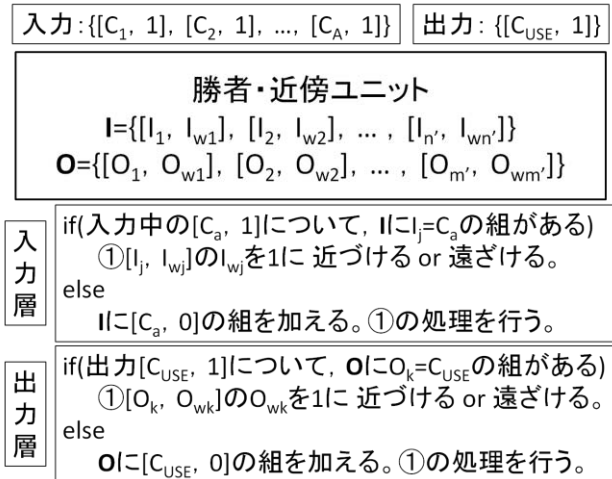


図 5 競合層の学習手順

学習後  $|\mathbf{I}| > n$  となった競合層ユニットは  $|\mathbf{I}| = n$  となるまで,  $\mathbf{I}$  の中で最も小さい重み  $I_{w_{min}}$  をもつ組を削除する.  $|\mathbf{O}| > m$  のユニットについても,  $\mathbf{O}$  の中で最小の重み  $O_{w_{min}}$  をもつ組を  $|\mathbf{O}| = m$  となるまで削除する(図 6).

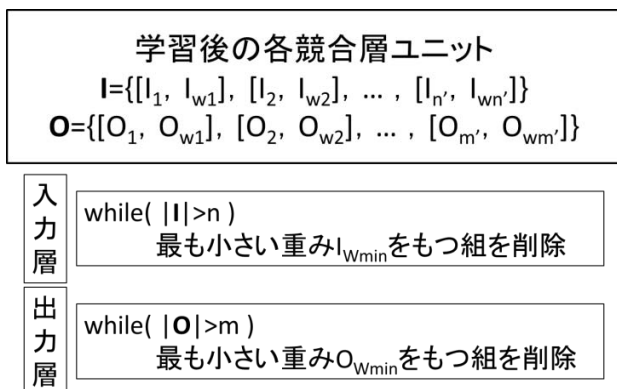


図 6 全学習終了後の処理

学習後の競合層ユニットは履歴コンテンツ集合  $\mathbf{I}_c = \{I_1, I_2, \dots, I_n\}$  と,  $\mathbf{I}_c$  の要素を利用履歴にもつユーザから高評価を受けた利用コンテンツ集合  $\mathbf{O}_c = \{O_1, O_2, \dots, O_m\}$  をもっている. また,  $\mathbf{I}_c$  の各要素の  $\mathbf{O}_c$  に対する高評価への関連度を表すベクトル  $\mathbf{I}_w = \{I_{w1}, I_{w2}, \dots, I_{wn}\}$  をもつ.

コンテンツベース型の場合と同様に, 学習後の競合層ユニットから図 3 のようなコンテンツマップを作成する. ただし, 既に出力情報がコンテンツの集合であるため, コンテンツ群を新たに構成する必要はない. 利用履歴  $\mathbf{L}$  をもつユーザに対する各競合層ユニットの類似度を式(7)から  $O_{w_c}$  の項を除いて求め, マップの表示に反映させる(図 4).

マップ操作はコンテンツベース型の場合と同様である. マップ非操作の場合, 式(5)に基づいて求めたベクトル中で値の大きな出力に対応したコンテンツを推薦する.

この手法では対象とした全ユーザの利用文脈, 評価と関連付けたコンテンツの重み付けを行っている. そのため, 学習後の競合層から全体傾向を抽出できる.

### 3. 検証実験

2.2 で述べたコンテンツベース型について実験を行った. 実験では Twitter で Yahoo!ニュース公式アカウントが配信している Web ニュースをコンテンツとし, Twitter 利用者をユーザとしてシステムを構築した. Twitter ユーザがニュースツイートをリツイートした事例を収集して学習データとした. 学習データの構造を表 1 に示す.

表 1 学習データの構造

	経済	芸能	スポーツ	IT	科学
入力(U)	$U_1$	$U_2$	$U_3$	$U_4$	$U_5$
出力(C)	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$

取得したテキスト中の全単語について 5 種の文書集合を適用した tf-idf 値[6]の合計を求め, 文書集合毎の合計値を 5 種全体の比率に直して学習用ベクトルとしている(図 7). 解析対象とする単語は形態素解析器 kuromoji[7]によって名詞かつ数でないものとされたものである. 解析に用いた文書集合に属する記事数と単語数を表 2 に示す. 記事は 2013/6/25~11/12 に Yahoo!ニュース RSS の各トピックで公開された中から収集した.

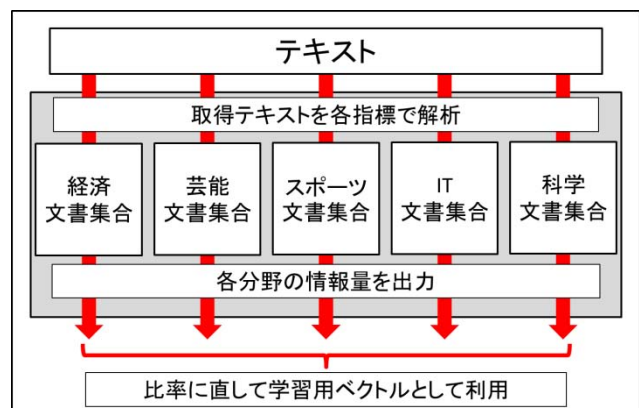


図 7 学習用ベクトルの生成

表 2 解析に用いた文書集合

	経済	芸能	スポーツ	IT	科学
記事数	198	222	225	166	167
単語数	4623	5691	3810	2212	3553

学習データは 2014/1/20~2014/1/29 の間に Yahoo!ニュース公式 Twitter アカウントから発信されたツイートの内,

20 人以上にリツイートされたニュースから生成した 210 件を用いている。各学習データは複数のユーザが同一のコンテンツを利用した事例集合であり、学習用ベクトルの集合  $\{(U_1, C), (U_2, C), \dots\}$  として扱われる。学習の際、リツイート行為は各ユーザのニュースに対する良い評価を表していると考えて、評価値  $E$  を 1 に固定して重みの更新を行っている。

3 パターンの実験を行い比較・考察した。全実験の共通事項として、210 件のデータ中の 7738 件のリツイート事例について 50 回のバッチ学習を終えた SORN を用いて推薦処理を行っている。学習パラメータの共通設定を表 3 に示す。 $\alpha(t)$  は各バッチ学習ごとに減少し、2 次元に配置した競合層について隣接関係から近傍ユニットを定めた。

表 3 学習パラメータの設定

$\alpha(t)$	$e^{-\frac{t+1}{50}} [1 \leq t \leq 50]$
$\beta$	なし(評価が負の学習を行わないため)
近傍ユニット	勝者ユニットの周囲 24 個について、 勝者の周りは更新値 0.5 倍、 その外周は更新値 0.2 倍として学習

提案手法であるコンテンツマップ方式(マップ方式)と総動員方式の特徴を図 8 に示す。マップ方式では単一ユニットが獲得した結合重みを利用するため、個々の事例に近いレベルの推薦を行う。総動員方式では全ユニットの学習結果の重み付き平均を利用するため、個人性のない全体傾向に従った推薦を行う。

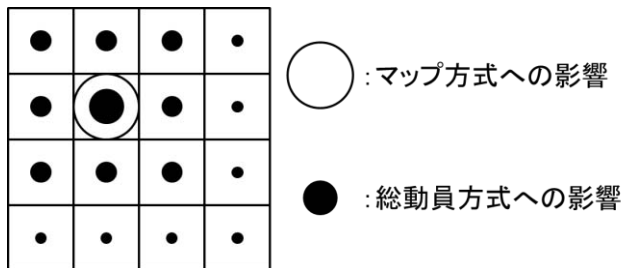


図 8 提案手法の特徴

### 3.1 コンテンツマップ方式(提案手法)

学習後の各競合層ユニットに 210 件のニュースから出力重みの近い上位数件のニュースを対応させる。その後、学習データ中の 7738 事例に含まれるユーザからランダムに 100 人抜き出し、各ユーザの属性に最も近い入力重みをもつ競合層ユニットを探索する。ユーザに最も近い重みをもつユニットに対応したニュースの中にユーザがリツイートしたものが含まれていれば推薦成功、含まれていなければ失敗としてカウントする。競合層のユニット数、1 ユニットに対応するニュース数を変えながら、各設定につき 100 回実験した平均の成功率を表 4 に示す。

表 4 ニュース推薦の成功率(マップ方式)

	ユニット数 10×10	ユニット数 15×15	ユニット数 20×20
対応数 5 件	12.5%	12.6%	12.8%
対応数 10 件	19.9%	19.2%	19.7%
対応数 15 件	24.7%	25.1%	25.4%

表 4 の行は 1 個の競合層ユニットに対応させるニュースの数が 5, 10, 15 件の設定に対応し、列は競合層ユニットを 10×10, 15×15, 20×20 の 2 次元に設定したことに対応する。ここで、表 4 中の成功率はコンテンツマップ上で最も強調されたマスにユーザが過去に利用したコンテンツが対応している確率である。推薦成功のケースでは過去に利用したコンテンツと共に、同系コンテンツの推薦に成功している。表 4 からはユニット数による推薦精度の変化は見られない。各設定 100 回の実験において競合層ユニットに対応しなかったニュース件数を表 5 に示す。

表 5 ユニット未対応のニュース件数

	ユニット数 10×10	ユニット数 15×15	ユニット数 20×20
対応数 5 件	145.4 件	99.9 件	75.1 件
対応数 10 件	115.6 件	72.3 件	52.5 件
対応数 15 件	99.7 件	56.8 件	39.4 件

表 5 より、210 件中かなりのニュースがマップに配置されていなかったことが分かる。ユニット数の増加に応じて未対応のニュースが減っていることから、競合層の出力重みの多様性はマップサイズに比例するといえる。

### 3.2 総動員方式(提案手法)

総動員方式では学習後の SORN にユーザ属性を与えて式(5)でニュース属性を求め、求めた属性に最も近い数件のニュースを推薦する。式(6)中の  $\gamma$  は 1 とした。マップ方式の実験と同様に、ユーザがリツイートしたニュースが推薦内容に含まれていれば成功、含まれていなければ失敗とした。学習データから無作為抽出した 100 人のユーザに対する実験を各設定 100 回行った平均成功率を表 6 に示す。

表 6 ニュース推薦の成功率(総動員方式)

	ユニット数 10×10	ユニット数 15×15	ユニット数 20×20
推薦数 5 件	9.9%	10.2%	10.4%
推薦数 10 件	20.0%	19.1%	19.7%
推薦数 15 件	25.4%	25.2%	25.3%

表 6 の行は各ユーザに対して推薦するニュースの数に対応し、列は競合層ユニットの設定に対応する。推薦数が 5 件のときマップ方式より精度が悪く、利用・被利用関係の平

均を求めて個人性を消した弊害が現れている。ユニット数による変化が見られないのはマップ方式と同じである。

### 3.3 直接推薦方式

学習データ中からランダムに抜き出した 100 人のユーザに対し、各ユーザ属性に最も近い特徴をもつニュース数件を直接推薦した。マップ方式、総動員方式と同様、推薦したニュースにユーザからリツイートされたものが含まれていれば成功とした。100 回実験した平均の成功率を表 7 に示す。この手法では学習を行わず、ユーザのツイート構成に近い話題のニュースを推薦している。

表 7 ニュース推薦の成功率(実験 3)

推薦数 5 件	14.8%
推薦数 10 件	23.8%
推薦数 15 件	30.3%

直接推薦方式ではマップ方式、総動員方式のいずれより高い精度を得た。これは、普段ツイートする内容に近い話題のニュースをリツイートするユーザが多いことを示す。学習データ中の 7738 事例について、ニュース属性とユーザ特徴の相関係数を求めた。表 8 に各次元ごとの相関係数を示す。各行がニュース特徴の次元、各列がユーザ属性の次元に対応しており、各セルにはそれぞれの行・列の値の相関係数が入っている。表 8 においてニュース・ユーザ間で同じ話題の次元同士が最も高い係数をもっている。相関係数から、普段のツイートと同じ分野のニュースをリツイートするユーザが多いことが確認できる。

表 8 ユーザ・ニュースの相関係数

	経 U	芸 U	ス U	IT U	科 U
経済 C	0.146	-0.060	-0.121	0.119	0.080
芸能 C	-0.110	0.172	-0.026	-0.074	-0.127
スポ C	-0.127	-0.028	0.218	-0.135	-0.093
IT C	0.092	-0.042	-0.104	0.148	0.047
科学 C	0.095	-0.089	-0.061	0.047	0.170

### 3.4 考察

データ中のユーザ 1 人が 210 件のニュースの中からリツイートしている平均件数は 1.58 件である(リツイート事例 7738 件、ユーザ 4905 人)。ランダムに推薦を行う場合の成功率を表 9 に示す。表 9 の値は式(8)で計算している。式(8)中の 210 はニュース件数、208 はニュース件数から平均リツイート数の切り上げ値を引いた数値である。提案手法 2 種はランダム推薦と比較すると良い結果だが、利用コンテンツの直接推定において有効な精度を得られたとは言えない。提案手法ではユーザとのインタラクションによる多様な推薦を想定しているため、マップ方式の勝者ユニット単独による推薦、総動員方式の平均化した 1 属性を用いた

推薦の単純な精度が評価の全てではないが、学習や推薦手法の改善が必要であると考えられる。

表 9 ランダム推薦の理論成功率

推薦件数	成功率
5 件	4.7%
10 件	9.3%
15 件	13.8%

$$\left\{ \begin{array}{l} \text{成功率} = 1 - \frac{\text{combin}(208, N)}{\text{combin}(210, N)} \\ N: \text{推薦件数} \end{array} \right. \quad (8)$$

マップサイズによる変化が見られない原因として、ニュース特徴の結合重みに対してユーザ属性の重みの多様性が低いことが考えられる。学習に用いたユーザ 7738 事例(4905 人の各リツイート時点)、ニュース 210 件の入出力ベクトルの各次元の平均・分散を表 10 に示す。

表 10 学習データの傾向

	経済	芸能	スポーツ	IT	科学
U の平均	0.2104	0.3156	0.2486	0.1263	0.0988
C の平均	0.2335	0.2506	0.2589	0.1102	0.1465
U の分散	0.0018	0.0053	0.0044	0.0010	0.0008
C の分散	0.0180	0.0256	0.0276	0.0076	0.0157

U・C 共に芸能・スポーツの平均が高く、Twitter 上のニュースや会話に芸能・スポーツの話題が多いことが分かる。これはネット上での芸能人やスポーツ選手の影響力の強さによると考えられる。また、C の分散に比べ U の分散は小さく、ユーザのツイートはニュースに比べて多様性が低いことが分かる。多様性の指標として、学習後の競合層にクラスタリングを行った。ユニット数を 20×20 として実験と同じ設定の学習を行なったユニットに k=20 の k-means 法[8]を適用した。クラスタリングには各ユニットの入力重みと出力重みの両方を用いている。得られたクラスタを図 9 に示す。図 9 中の上の棒グラフがニュース特徴、下がユーザ属性を表す。各グラフの左から経済、芸能、スポーツ、IT、科学の割合に対応する。クラスタ番号隣の括弧内の数字は所属ユニット数である。なお、空のクラスタは載せていない。得られたクラスタでは下のユーザ属性に大きな差は見られないが、ニュースの特徴には明確な違いが存在する。また、ニュース特徴・ユーザ属性の重みを別々にクラスタリングした場合でも同様の傾向が現れている。ユーザ多様性が低いと各ユニット間でのニュース特徴の差分が相対的に大きくなり、ユーザ属性に過敏な推薦を行って推薦内容が分散していると考えられる。ユーザやコンテンツの学習データに統計的な差異がある場合に近似度計算の重み付け配分を変えるなど、両者の均衡をとる仕組みにより推薦精度が改善すると予想できる。また、ユーザ・コンテ

コンテンツの解析自体に多様な学習データを生成する仕組みを取り入れることも有効と考えられる。

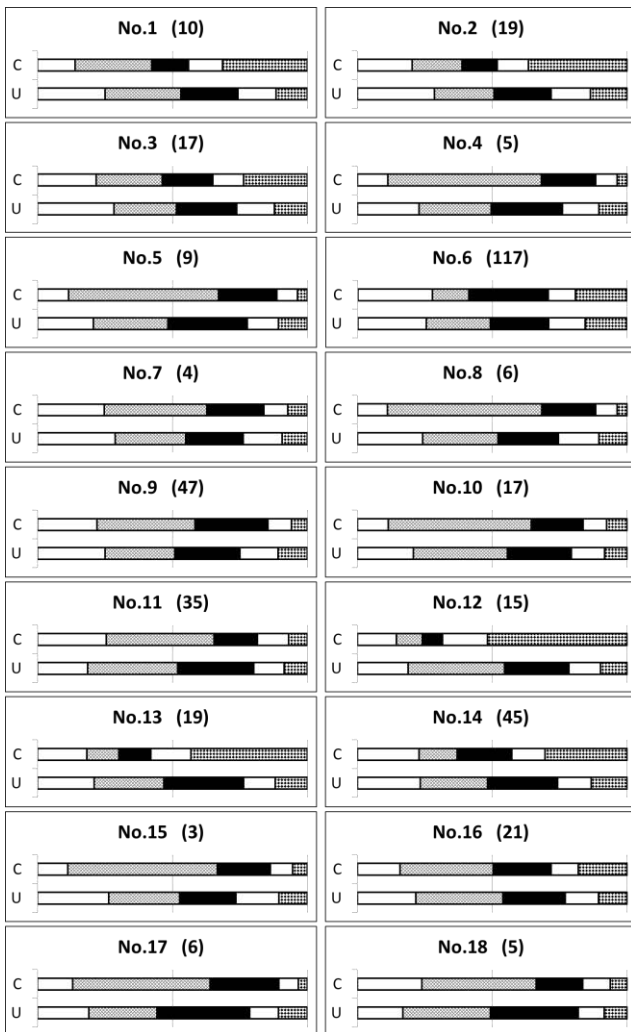


図9 競合層のクラスター

直接推薦方式との比較から、ユーザとコンテンツのパラメータ間に明確な相関関係が含まれる場合、提案手法は非学習の最尤推定手法の推薦に対して精度が劣ると予想される。図9に示される学習結果は、似た嗜好のユーザが異なる分野のニュースを評価した事例を獲得した結果だと判断できる。この学習結果は、提案手法でのマップ操作による推薦を行うための学習に成功していると言える。マップ方式の実験では推薦に用いるユニットをユーザ属性に最も近いものとしたが、本来はユーザに選択させるものである。クラスタリングの結果を踏まえると、ユーザに推薦内容を選択させる方式であれば単純な精度以上の評価を獲得できると考えられる。

マップ方式と総動員方式を比較すると、マップ方式では個々の事例に近い推薦、総動員方式では個性を排した推薦を行っている。両手法の違いは利用する競合層ユニットが単一か複数かであり、利用ユニットの決め方や数の設定で推薦の性質を調整できる。クラスタリング後のラベルや

マップ中のユニット位置情報などがユニット選定に利用可能である。各指標で選定されたユニット群による協調推薦では、特定ユーザ層の利用コンテンツ群を辿る階層移動的な推薦が可能になると考えられる。

#### 4. まとめ

本論文ではユーザとコンテンツ双方の特徴を結びつけた機械学習による推薦システムを提案した。提案システムはコンテンツベース型、協調フィルタリング型の推薦に適用できる。また、グラフィカルなマップにコンテンツを配置したユーザ操作を伴うインタラクティブな推薦、学習結果全体の傾向を抽出した推薦が可能である。コンテンツベース型の実験において良い推薦精度を得られなかったが、マップ操作により他者の多様な興味を反映させるインタラクティブな推薦を行える可能性が示された。

提案手法による有効な推薦にはユーザ多様性とコンテンツ多様性のバランスが重要と考えられる。ユーザ側重みの多様性が欠けるとインタラクティブな推薦の意義が失われ、コンテンツ側重みの多様性がないと単調な推薦しか行えない。多数派の性質を強く学習すると多様性が失われ、少数派を重視すると推薦の根拠が弱くなる。有効なバランスを保つためには、ユニークなユーザ・コンテンツの利用関係を適切に競合層に保持しながら大勢を獲得する学習機構が必要である。学習パラメータの設定によっても推薦の性質が変わると考えられるため、有効な推薦に関わる要素を明らかにしながら適切な学習手法を求めていく必要がある。

今後は多様性を確保するための学習手法を考えながら、ハイパーパラメータの影響や協調フィルタリング方式の性質を測るための実験を行っていく予定である。

#### 参考文献

- [1]麻生 英樹, 小野 智弘, 本村 陽一, 黒川 茂莉, 櫻井 彰人, “協調フィルタリングと属性ベースフィルタリングの統合について”, 電子情報通信学会技術研究報告. NC, Vol.106, No.279, pp.55-59 (2006)
- [2]土方 嘉徳, “嗜好抽出と情報推薦技術”, 情報処理学会誌, Vol.48, No.9, pp.957-965 (2007)
- [3] G. Adomavicius, A. Tuzhilin, “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions”, IEEE Trans. on Knowledge and Data Eng, Vol.17, No.6, pp.734-749 (2005)
- [4] 堀尾 恵一, 原口 琢磨, 山川 烈, “自己組織化関係ネットワークを用いた主観に基づく画像変換”, バイオメディカル・ファジィ・システム学会誌, Vol.1, No.1 pp. 79-86 (1999)
- [5] Kohonen, Teuvo, “The self-organizing map”, Proceedings of the IEEE, Vol.78, No.9, pp.1464-1480 (1990)
- [6] 仲尾由雄, “文書の話題構成に基づく重要語の抽出”, 情報処理学会研究報告. 情報学基礎研究会報告, Vol.98, No.60, pp.1-8 (1998)
- [7] atilika applied search innovation, kuromoji-0.7.7: <http://www.atilika.com/ja/products/kuromoji.html>
- [8] 神島敏弘, “データマイニング分野のクラスタリング手法(1)”, 人工知能学会誌, Vol.18, No.1, pp.59-65 (2003)