RUI-Filtering:利用履歴のアイテムの類似関係を反映した協調フィルタリング方式

土井 俊介　　吉田 由紀　　東野 豪

日本電信電話株式会社　ＮＴＴサイバーソリューション研究所
〒239-0847 神奈川県横須賀市光の丘 1-1
E-mail:　{doi.shunsuke, yoshida.yuki, higashino.suguru}@lab.ntt.co.jp

あらまし　Content-based フィルタリングにおいて，他ユーザの評価の反映を可能とすることを目的とするフィルタリング方式を提案する．複数ユーザのコンテンツ利用履歴から，類似関係があるコンテンツのメタ情報のグループを生成し，そのグループを介してフィルタリングを行った．その結果，提案方式は Content-based フィルタリングと同等の正答率であったが，Content-based フィルタリングでは推薦できなかったコンテンツを約 5%選出できた．また推薦コンテンツ数を絞った場合において，Content-based フィルタリングより高い性能が得られた．

キーワード　個人化サービス, 協調フィルタリング，Content-based フィルタリング, 推薦システム

RUI-Filtering: A content-based, collaborative filtering algorithm using the access log of related items

Shunsuke DOI　　Yuki YOSHIDA　and　Suguru HIGASHINO

NTT Cyber Solutions Laboratories, NTT Corporation　1-1, Hikarinooka Yokosuka-Shi Kanagawa, 239-0847 Japan
E-mail:　{doi.shunsuke, yoshida.yuki, higashino.suguru}@lab.ntt.co.jp

Abstract　For content-based filtering, we propose a new filtering method that can reflect the evaluation of another user. Some groups of meta-data are generated among similar contents from many users' access logs, whereas other contents are filtered by group. The method was evaluated to be equivalent to the traditional content-based filtering, and it could recommend contents that were not recommended by the traditional method about 5% of the time. Moreover, when we restricted the number of recommendation contents, our method showed higher performance than the traditional method.

Keyword　Personalized Service, Collaborative filtering, Content-based filtering, Recommendation System

1. はじめに

ネットワークから多くのコンテンツが利用できる環境が整いつつある今日，多くのコンテンツの中からユーザが希望とするコンテンツを発見しやすい環境を提供することが求められている．

ユーザの好みに応じてコンテンツを選出する方法として典型的な 2 つの方法がある．1 つは Content-based フィルタリング（内容に基づくフィルタリング）と，もう 1 つは協調フィルタリングである．

Content-based フィルタリングは，あるユーザに対し，そのユーザが過去に評価を行ったコンテンツの内容を分析して得られたユーザプロファイルに基づいて推薦される[1]．つまり，未知のコンテンツであっても，ユーザプロファイルとコンテンツの内容が類似していれば推薦できる．しかし，過去の評価が不十分な場合，満足にコンテンツを推薦できず，評価内容に類似したコンテンツばかりが推薦される傾向がある．

一方，協調フィルタリングは，類似した好みを有する他のユーザが好んだコンテンツを推薦する[2,3,4]．コンテンツの内容についての評価は行わないため，そこに付与された情報（メタ情報）は不要である．更に，他のユーザの評価を反映した推薦ができ，ユーザの過去の評価が少ないため Content-based フィルタリングでは類似性が判別できず推薦されないコンテンツも推薦できる．しかし，他のユーザが未評価の新しいコンテンツは推薦できない．

そこで，Content-based フィルタリングに協調フィルタリングの効果を得ることができれば，未評価の新しいコンテンツを推薦する際に，協調フィルタリングの効果をもたらすことができると考える．例えば，新製品のダイレクトメールを，新製品の情報と事前に登録された嗜好情報とが類似するユーザを対象として一斉

に送信する場合，本当は興味があるのに送信対象にならなかったユーザを発見することが可能となる．

本稿では，Content-based フィルタリングにおいて，他ユーザの評価の反映をする方式を提案し，その方式の手順と評価結果について述べる．

2. RUI-Filtering の提案

2.1. 問題意識

本研究では，Content-basedフィルタリングにおいて，他ユーザの評価を反映することで，ユーザに適しているにもかかわらず Content-based フィルタリングでは推薦できないコンテンツの推薦を可能とすることを目的とする．図 1 に本研究の対象範囲を示す．

	他ユーザの評価を反映する	他ユーザの評価を反映しない
未評価コンテンツの推薦が可能，コンテンツにメタ情報が必要	**対象範囲**	Content-based フィルタリング
未評価コンテンツの推薦が不可能，コンテンツにメタ情報は不要	協調フィルタリング	

図 1:本研究の対象範囲

2.2. 関連研究

Content-based フィルタリング，協調フィルタリングそれぞれの弱点を解決するために，両者を組み合わせた研究は行われている．

Fab は，WWW のページを収集提示するシステムである．ユーザは自分のプロファイルで高く評価される情報と，類似するユーザが高く評価する情報の両方を受けとる[5].

浅川らは，ユーザの嗜好の部分的な類似性を利用した協調フィルタリングを提案し，ジャンルといった部分的に嗜好が類似したユーザグループを生成して協調フィルタリングに応用している[6].

また，利用履歴から生成したカテゴリを応用する例として，岡本らは利用履歴から類似したコンテンツをカテゴリに分け，そのカテゴリ別にコンテンツ推薦を行っている[7].

その他に，利用のされ方が類似したコンテンツに着目した Item-based の協調フィルタリング方式も提案されている[8].

2.3. 仮説

筆者らは，これまでにテレビ番組の視聴履歴から，視聴した番組のキーワード（出演者・タイトル・放送局といったメタ情報）に視聴時間や視聴回数で重み付けをし，クラスタリングを行ってキーワードのグループを複数個生成し，それぞれのグループを調べたところ，同一のグループに含まれるキーワードは，複数ユーザに渡った視聴の特徴を表す傾向がみられる，という結果を得た[9].

この傾向を応用して「複数ユーザのコンテンツの利用履歴や評価結果からしかるべき方法によって生成した複数個のキーワードグループを，コンテンツ推薦に応用することで，複数の他ユーザの利用履歴から現れる特徴を反映した協調的なコンテンツ推薦が可能になる」との仮説を立てる．

2.4. キーワードグループを介した類似度算出

前記の仮説に基づき，キーワードグループを介してユーザとコンテンツ間の類似度を算出する方式を提案する．図 2 にその模式図を示す．

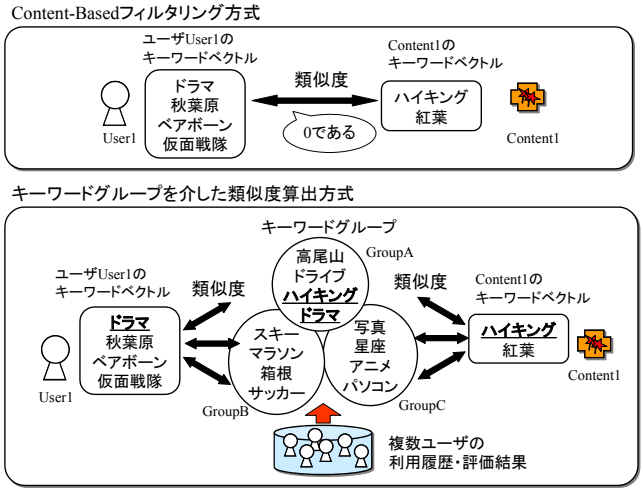


図 2:キーワードグループを介した類似度算出

図 2 では，ユーザ *User1* のキーワードベクトル $\boldsymbol{U}_{User1}$={ドラマ，秋葉原，ベアボーン，仮面戦隊}，コンテンツ *Content1* のキーワードベクトル $\boldsymbol{C}_{Content1}$={ハイキング，紅葉} となっている．また，キーワードグループは，*GroupA*={高尾山，ドライブ，ハイキング，ドラマ}，*GroupB*={スキー，マラソン，箱根，サッカー}，*GroupC*={写真，星座，アニメ，パソコン} となっている．キーワードグループは，複数ユーザの利用履歴から既に生成されたものである．

図 2 に示すように Content-based フィルタリングでは，$\boldsymbol{U}_{User1}$ と $\boldsymbol{C}_{Content1}$ との類似度はキーワードベクトル同士が独立しているため 0 となり推薦されない．

しかし，キーワードグループを介して類似度を算出する方式の場合，キーワードグループ *GroupA* にある「ドラマ」によって *GroupA* と *User1* のキーワードベクトル間に類似度が得られる．また，キーワードグループ *GroupA* にある「ハイキング」によって *GroupA* と *Content1* のキーワードベクトル間に類似度が得られ，値の大きさによっては推薦も可能となる．

つまり Content-based フィルタリングでは，$\boldsymbol{U}_{User1}$ と $\boldsymbol{C}_{Content1}$ との間に類似度は算出されなかったが，キーワードグループ *GroupA* に「ドラマ」と「ハイキング」が含まれていたため，このキーワードグループ *GroupA*

を介すると類似度の算出が可能となった.

本稿ではこの方式を，キーワードグループとの類似度を介してユーザ・コンテンツ間の類似度を算出し，コンテンツの推薦に用いることから「類似度」にちなんで RUI-Filtering と呼び，提案する.

なお，本稿において，キーワードとはコンテンツに付与されたメタ情報，キーワードグループとはメタ情報の集合であると定義する.

2.5. RUI-Filtering によるコンテンツ推薦手順

以下の手順によって RUI-Filtering を用いたコンテンツ推薦を行う.

- **Step1.キーワードベクトルの生成**
- **Step2.キーワードグループの生成**
- **Step3.RUI-Filtering による類似度の算出**
- **Step4.類似度の大きさによる推薦の判別**

Step1.キーワードベクトルの生成

キーワードベクトルは，キーワードに対して重み値が付与されたベクトルである．各ユーザのキーワードベクトル（ユーザ u のキーワードベクトル $\boldsymbol{U}$）を式(1)のように表す.

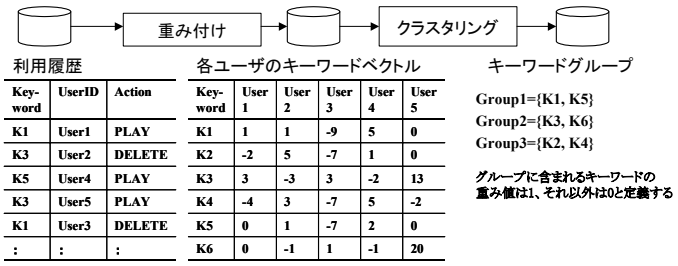
$$\boldsymbol{U} = (w_{uK1}, w_{uK2}, ..., w_{uKn}) \quad (1)$$

ここで w_{uK1} はユーザ u のキーワード $K1$ の重み値をあらわす．n はキーワードの総数である.

本稿では，各ユーザのキーワードベクトルの生成方法についての定義は行わないが，以下に，コンテンツの利用履歴から各ユーザのキーワードベクトルを生成する例を示す.

コンテンツには，それぞれコンテンツの内容を表すメタ情報として，キーワードがあらかじめ付与されているとする．キーワードはコンテンツのタイトルや出演者,製作者や説明文を形態素解析した単語等である.利用履歴は，コンテンツを利用したユーザ ID と，利用したコンテンツのキーワードと，そのコンテンツに対する操作情報が記録されているとする.

そこで，ユーザがコンテンツに対して行った操作情報に応じてキーワードに重み付けを行うことで，キーワードベクトルを生成する．図 3 に各ユーザのキーワードベクトルの生成とキーワードグループの生成例を示す．例えば，コンテンツ利用履歴において *User1* がキーワード *K1* を含むコンテンツを”PLAY”すれば，*User1* の *K1* の重み値を+1 し，*User2* が *K3* を含むコンテンツを”DELETE”すれば *User2* の *K3* の重み値を-1 する，というようにあらかじめ定義したルールに従って重み付けを行う．図 3 の *User1* のキーワードベクトル $\boldsymbol{U}_{user1}$ は，$\boldsymbol{U}_{user1}$=(1,-2, 3,-4, 0, 0) となる.



Key-word	UserID	Action
K1	User1	PLAY
K3	User2	DELETE
K5	User4	PLAY
K3	User5	PLAY
K1	User3	DELETE
:	:	:

Key-word	User 1	User 2	User 3	User 4	User 5
K1	1	1	-9	5	0
K2	-2	5	-7	1	0
K3	3	-3	3	-2	13
K4	-4	3	-7	5	-2
K5	0	1	-7	2	0
K6	0	-1	1	-1	20

図 3:各ユーザのキーワードベクトル，
キーワードグループの生成例

Step2.キーワードグループの生成

Step1.で生成した各ユーザのキーワードベクトルを用いて，キーワードグループを生成する.

キーワードグループの生成には，任意の数のクラスタを高速に生成できることから，K-Means 法によるクラスタリングを用いる．この際，複数ユーザのキーワードベクトルを入力として，生成するキーワードグループの数（クラスタ数）を指定し，キーワードをクラスタリングする．生成したキーワードグループ g_j のキーワードベクトル $\boldsymbol{G}_j$ は，式(2)のように表す.

$$\boldsymbol{G}_j = (w_{gjK1}, w_{gjK2}, ..., w_{gjKn}) \quad (2)$$

ここで，w_{gjK1} はキーワードグループ g_j のキーワード $K1$ の重み値をあらわす．図 3 のキーワードグループ *Group1* のキーワードベクトルは，*Group1* に含まれるキーワードが *K1*,*K5* で，キーワードグループに含まれるキーワードの重み値は 1，それ以外は 0 と定義されているので，$\boldsymbol{G}_{group1}$=(1, 0, 0, 0, 1, 0) となる.

なお，K-Means 法によるクラスタリングによるキーワードグループの生成手法以外にも，異なるクラスタリング手法による場合や，データマイニングの相関ルール分析によって相関関係が抽出されたキーワード同士をグループ化してキーワードグループを生成する方法等も考えられる[9].

Step3. RUI-Filtering による類似度の算出

RUI-Filtering では，ユーザのキーワードベクトルと，推薦対象となるコンテンツのキーワードベクトル，さらに，Step2.で生成した m 個のキーワードグループのキーワードベクトルを用いる.

推薦対象となるコンテンツ c のキーワードベクトル $\boldsymbol{C}$ は式(3)のように表す.

$$\boldsymbol{C} = (w_{cK1}, w_{cK2}, ..., w_{cKn}) \quad (3)$$

ここで，w_{cK1} はコンテンツ c のキーワード $K1$ の重み値をあらわす．n はキーワードの総数である．例えば，コンテンツ *Content1* に付与されたキーワードが *K5*,*K6* で，付与されたキーワードの重み値がそれぞれ，5,6 の場合，$\boldsymbol{C}_{Content1}$=(0, 0, 0, 0, 5, 6) となる.

次に，RUI-Filtering によって類似度を算出する．RUI-Filtering によるユーザ u のキーワードベクトル $\boldsymbol{U}$ と，コンテンツ c のキーワードベクトル $\boldsymbol{C}$ 間の類似度 $Rui(\boldsymbol{U},\boldsymbol{C})$ の算出式を式(4)に示す．

$$Rui(\boldsymbol{U},\boldsymbol{C}) = \sum_{j=1}^{m} Sim(\boldsymbol{U},\boldsymbol{G}_j) \cdot Sim(\boldsymbol{C},\boldsymbol{G}_j) \qquad (4)$$

ここで m はキーワードグループの総数である．$Sim(\boldsymbol{U},\boldsymbol{G}_j)$ はユーザ u のキーワードベクトル $\boldsymbol{U}$ とキーワードグループ g_j のキーワードベクトル $\boldsymbol{G}_j$ の類似度を，$Sim(\boldsymbol{C},\boldsymbol{G}_j)$ はコンテンツ c のキーワードベクトル $\boldsymbol{C}$ とキーワードグループ g_j のキーワードベクトル $\boldsymbol{G}_j$ の類似度を表す．類似度の算出法として，余弦，内積，Dice 係数，Jaccard 係数，相関係数などを用いることができる[10]．本稿では，ベクトル間の角度で類似度を算出し，ベクトルの大きさの影響を受けず，計算量の少ない余弦を用いる．

余弦による類似度算出式 $Sim(\boldsymbol{U},\boldsymbol{G}_j)$, $Sim(\boldsymbol{C},\boldsymbol{G}_j)$ を式(5)に示す．n はキーワードの総数である．

$$\left.\begin{aligned} Sim(\boldsymbol{U},\boldsymbol{G}_j) &= \frac{\sum_{i=1}^{n} w_{uK_i} \cdot w_{g_jK_i}}{\sqrt{\sum_{i=1}^{n} w_{uK_i}^{\ 2} \times \sum_{i=1}^{n} w_{g_jK_i}^{\ 2}}} \\ Sim(\boldsymbol{C},\boldsymbol{G}_j) &= \frac{\sum_{i=1}^{n} w_{cK_i} \cdot w_{g_jK_i}}{\sqrt{\sum_{i=1}^{n} w_{cK_i}^{\ 2} \times \sum_{i=1}^{n} w_{g_jK_i}^{\ 2}}} \end{aligned}\right\} \qquad (5)$$

次に，図 4 を用いて RUI-Filtering による類似度の算出例を説明する．

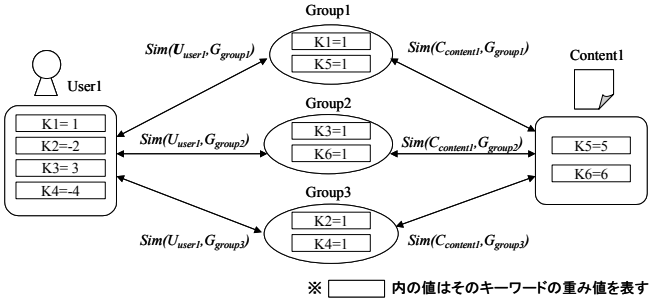


図 4：キーワードグループを介した類似度の算出例

図 4 において，それぞれのキーワードベクトルは，

$$\boldsymbol{U}_{User1} = (1,-2,3,-4,0,0)$$
$$\boldsymbol{C}_{content1} = (0,0,0,0,5,6)$$
$$\boldsymbol{G}_{group1} = (1,0,0,0,1,0)$$
$$\boldsymbol{G}_{group2} = (0,0,1,0,0,1)$$
$$\boldsymbol{G}_{group3} = (0,1,0,1,0,0)$$

となっている．ここで式(4)(5)を用いて RUI-Filtering による *User1* と *Content1* 間の類似度 $Rui(\boldsymbol{U}_{User1},\boldsymbol{C}_{content1})$ を算出すると，

$$\begin{aligned} &Rui(\boldsymbol{U}_{User1},\boldsymbol{C}_{content1}) \\ &= Sim(\boldsymbol{U}_{User1},\boldsymbol{G}_{group1}) \cdot Sim(\boldsymbol{C}_{content1},\boldsymbol{G}_{group1}) \\ &\quad + Sim(\boldsymbol{U}_{User1},\boldsymbol{G}_{group2}) \cdot Sim(\boldsymbol{C}_{content1},\boldsymbol{G}_{group2}) \\ &\quad + Sim(\boldsymbol{U}_{User1},\boldsymbol{G}_{group3}) \cdot Sim(\boldsymbol{C}_{content1},\boldsymbol{G}_{group3}) \\ &= (0.129 \cdot 0.453) + (0.387 \cdot 0.543) + (-0.775 \cdot 0.0) \\ &\fallingdotseq 0.269 \end{aligned}$$

となる．なお，余弦や内積を用いて *User1* と *Content1* のキーワードベクトル同士の類似度を直接算出した場合，0 となる．

Step4. 類似度の大きさによる推薦の判別

上記 Step.3 で算出した類似度 $Rui(\boldsymbol{U},\boldsymbol{C})$ を全てのユーザ・コンテンツに対して算出し，コンテンツ c をユーザ u に推薦するか否かのフィルタリングを行う．例えば，あるしきい値を設定し，そのしきい値を上回っているならば推薦する，しきい値以下ならば推薦しないとする．

3. 実験

被験者によるコンテンツ情報選択実験を行い，提案手法によるコンテンツフィルタリングの評価を行った．

3.1. コンテンツ選択実験

実験日から約半月後に神奈川・東京において実際に行われるイベント情報（テキストデータのみ）をコンテンツとして用い，Web ブラウザによって被験者 26 名に提示する．それを閲覧した被験者はそのイベント情報に対して「保存」「削除」のいずれかの選択を行う．「保存」「削除」のいずれかを選択の後は，引き続き次のイベント情報が提示される．イベント情報は合計 100 個提示される．全ての被験者に対して同じ順序で同じ内容のイベント情報が提示される．Web ブラウザによってコンテンツを提示している画面例を図 5 に示す．実験で用いたコンテンツに関する情報を表 1 に示す．また，被験者の選択結果を表 2 に示す．

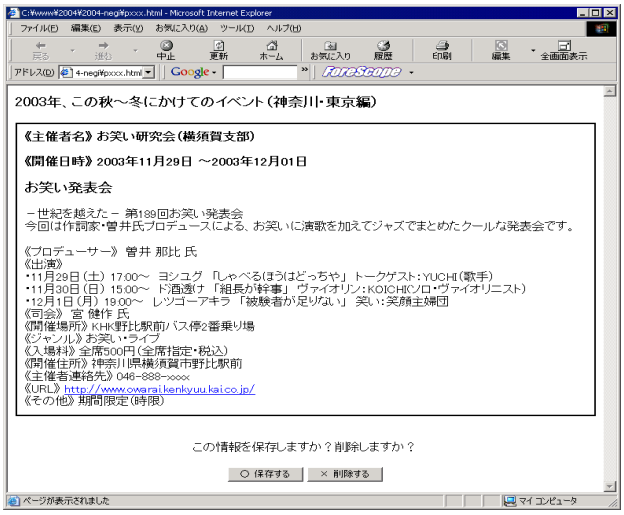


図 5：コンテンツ選択実験画面例

表 1:評価実験で用いたコンテンツ

コンテンツ内容	テキストデータ（イベント情報）
提示コンテンツ数	100 個
コンテンツのキーワード（メタ情報）の付与方法	コンテンツ(テキストデータ)を形態素解析し，文字長 2 以上の名詞と未知語をキーワードとする
コンテンツに付与されたキーワード（メタ情報）数	1 コンテンツあたり平均 129.8 個 キーワードを重複排除した場合 前半 50 コンテンツ =2043 個 後半 50 コンテンツ =2067 個 前半・後半重複キーワード=814 個
形態素解析ツール	茶筌 ver.2.1[11]

表 2:コンテンツ選択実験結果

被験者数	26 名
選択の割合	「保存」平均 27.2% 「削除」平均 72.8%

3.2. 従来方式と提案方式による選択の予測

次の手順によって，従来方式と提案方式によるコンテンツの選択内容を予測する．

①26 名の被験者から得られた 100 コンテンツに対するコンテンツ選択実験結果を前後の 2 つに分け，前半コンテンツの選択結果を「学習データ」，後半コンテンツの選択結果を「正解データ」とする．

②「学習データ」から全ユーザのキーワードベクトルとキーワードグループを生成する．

③各ユーザのキーワードベクトル，キーワードグループ，後半コンテンツのキーワードベクトル（メタ情報）を用いて，後半コンテンツの選択内容を予測（コンテンツの推薦の可否を判別）する．

④「正解データ」と予測内容とを比較し，一致したものを「正解」，不一致を「不正解」として正答率を算出する．

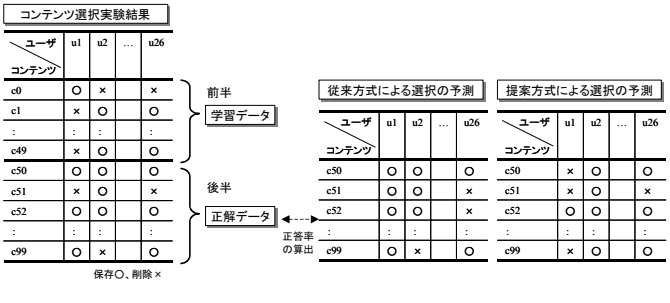


図 6:選択内容の予測と比較

(1) キーワードベクトルの生成

ユーザ u が「保存」を選択した場合，そのコンテンツに含まれるキーワード K のスコア S_{uK} を 1 加算する．「削除」を選択した場合，同様に 1 減算する．キーワード重み付けが完了した後，式(6)によってキーワード K 毎に正規化を行う．このようにして各ユーザのキーワードベクトル $\boldsymbol{U}=(w_{uK1},w_{uK2,\ldots,}w_{uKn})$ を生成する．n はキーワードの総数である．

$$w_{uK_i}=\frac{S_{uK_i}}{\sum_{i=1}^{n}\left|S_{uK_i}\right|} \tag{6}$$

生成した各ユーザのキーワードベクトルの例を表 3 に示す．

表 3:生成した各ユーザのキーワードベクトル例（一部抜粋）

キーワード	u1	u2	u3	u4	u5	u6	…	u26
CAFE	−0.038	−0.038	−0.038	−0.038	−0.038	−0.038		−0.038
サルサ	−0.038	−0.038	−0.038	−0.038	−0.038	−0.038		−0.038
レゲエ	−0.038	−0.038	−0.038	−0.038	−0.038	−0.038		−0.038
ソウル	−0.038	−0.038	−0.038	−0.038	−0.038	−0.038		−0.038
ライブ	−0.037	−0.037	−0.024	−0.037	−0.024	−0.049		−0.037
連絡	−0.004	−0.007	0.004	−0.032	−0.021	−0.057		−0.071
交通	0.000	−0.007	0.007	−0.029	−0.026	−0.051		−0.070
手段	−0.004	−0.004	0.012	−0.028	−0.024	−0.049		−0.069
YRP	−0.038	−0.038	−0.038	−0.038	−0.038	−0.038		−0.038
：								
野比	−0.038	−0.038	−0.038	−0.038	−0.038	−0.038		−0.038
駐車	−0.037	−0.022	−0.022	−0.022	−0.051	−0.037		−0.051
ランド	0.000	0.108	0.081	0.000	0.000	−0.108		0.000

コンテンツ c のキーワードベクトル $\boldsymbol{C}$ $\boldsymbol{C}=(w_{cK1,}w_{cK2,\ldots,}w_{cKn})$ は，コンテンツに付与されたキーワード K の重み w_{cK} を 1 とし，それ以外を 0 とする．

(2)キーワードグループの生成

上記(1)で生成した全ユーザのキーワードベクトルを K-Means 法でクラスタリング(正規化)してキーワードグループ g の生成を行う．ここでは，生成するキーワードグループの数による違いを検証するため 5，10，15，20，25，30 個の計 6 パターンのキーワードグループを生成する．多くのキーワードグループを生成すれば，一つのキーワードグループに含まれるキーワードの数は少なくなる．生成したキーワードグループの例を表 4 に示す．キーワードグループ g のキーワードベクトル $\boldsymbol{G}$ は，キーワードグループ g に存在するキーワード K の重み値 w_{gK} を 1 とし，存在しない場合は 0 とする．

表 4:生成したキーワードグループ例（一部抜粋）

Group1	Group2	Group3	Group4	Group5	Group6	…
コンサート	マーク	クリスマスツリー	水面	土曜	ボーラ	
テーマ	白夜	巨大	首都	財団	振興	
ベル	イメージ	輝き	ボード	エコロジー	会期	
史上	幻想	サンタ	ウォーク	企画	無休	
火曜	過去	ペンギン	直接	特別	IMAGINE	
丸の内	12M	パレード	イルカ	先駆	コレクション	
ビジネスマン	スクエア	トナカイ	泳ぎ	紹介	エポック	
実践	ろうそく	変身	ジャンプ	展示	箱根登山鉄道	
中心	オーナメント	ケープペンギン	間近	いろいろ	前後	
天国	精霊	ペタペタ	楽園	最新	芸術	
地獄	吹き抜け	チョコ	エサ	都心	印象派	
天使	圧倒的	フローティング	午前	多数	象徴	
悪魔	ハート	8M	高校生	挑戦	橋	
恐怖	それ	湾内	ゲリラ	生活	アール	
オーロラ	天秤	観覧	京王	視点	ヌーヴォー	
ライティング	景品	セレモニー	小田急	会場	ガラス	
：						

(3) RUI-Filtering と従来方式による類似度の算出

提案方式である RUI-Filtering によるユーザ・コンテンツ間類似度は式(4)を用いて算出する．実験結果から，全 26 ユーザの前半 50 コンテンツを「学習データ」，後半 50 コンテンツを「正解データ」とした場合，ユーザ $u1$～$u26$ のキーワードベクトル $\boldsymbol{U1}$～$\boldsymbol{U26}$ と後半 50 コンテンツ $c50$～$c99$ のキーワードベクトル $\boldsymbol{C50}$～$\boldsymbol{C99}$ との各類似度 $Rui(\boldsymbol{U1},\boldsymbol{C50})$, $Rui(\boldsymbol{U1},\boldsymbol{C51})$, …, $Rui(\boldsymbol{U26},\boldsymbol{C99})$

（計 1300 個の類似度）を算出する．式(4)における $Sim(U,G_j)$，$Sim(C,G_j)$ は式(5)に示す余弦を用いる．

従来方式である Content-based によるフィルタリングは，ユーザ u のキーワードベクトル $\boldsymbol{U}$ とコンテンツ c のキーワードベクトル $\boldsymbol{C}$ との類似度を算出し，その大きさによってフィルタリングを行う．ここでは類似度算出に余弦を用いる．余弦を用いた類似度算出式を式(7)に示す．

$$Sim(\boldsymbol{U},\boldsymbol{C}) = \frac{\sum_{i=1}^{n} w_{uK_i} \cdot w_{cK_i}}{\sqrt{\sum_{i=1}^{n} w_{uK_i}^{\ 2} \times \sum_{i=1}^{n} w_{cK_i}^{\ 2}}} \qquad (7)$$

式(7)を用いて，従来方式によるユーザ・コンテンツ間類似度 $Sim(\boldsymbol{U}1,\boldsymbol{C}50)$，$Sim(\boldsymbol{U}1,\boldsymbol{C}51),\ldots,Sim(\boldsymbol{U}26,\boldsymbol{C}99)$ を算出する．

(4) 類似度の大きさによる推薦の判別

算出した類似度の大きさから，しきい値によって「保存」か「削除」かを判別する．本評価では最適なしきい値を用いるために，しきい値を低い値から高い値へと 0.00001 刻みに変化させ，各しきい値において判別結果と「正解データ」とを比較し，正答率が最大になった時点のしきい値を最適なしきい値として用いた．しきい値は，前記(3)で生成した各ユーザ・各コンテンツ間の類似度（26 ユーザ×50 コンテンツの場合，1300 個の類似度）に対して同一の値を用いて判別した．

ただし，学習データや正解データを変更した場合や，使用するキーワードグループを変更した場合は，再度最適なしきい値を求めた．また従来方式においても別途最適なしきい値を設定した．

例えば，提案方式において，「学習データ」を前半 50 コンテンツ，「正解データ」を後半 50 コンテンツ，5 個生成したキーワードグループを用いた場合，最適なしきい値は 0.00003 であった．同条件の場合，従来方式の最適なしきい値は，-0.00151 であった．

3.3 評価

3.3.1 正答率による評価

(1)キーワードグループ数による正答率の変化

図 7 に提案方式と従来方式の正答率を比較したグラフを示す．提案方式ならびに従来方式で予測した全 26 ユーザの後半 50 コンテンツの予測内容と「正解データ」とを比較し，一致したものを「正解」として正答率を算出した．図 9 の破線は全てを「削除」と予測した場合の正答率のラインを示す．提案方式は平均 80.64%，従来方式は平均 80.77%の正答率であった．用いたキーワードグループによって正答率に変動は生じたが，有意な差ではなかった．つまり，両方式はほぼ同等の正答率となった．

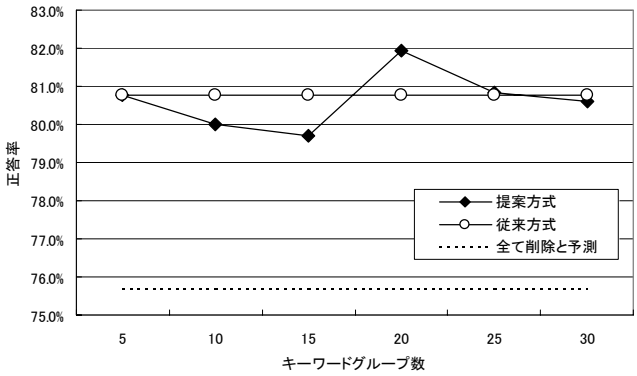


図 7:キーワードグループ数による正答率の変化

(2)提案方式と従来方式との差異

提案方式と従来方式において予測内容が相反していた個所に着目し，相反していた個所と実際コンテンツの選択結果を比較し，提案方式が正解(従来方式が不正解)の割合と提案方式が不正解(提案方式が正解)の割合を算出した．図 8 にその結果を示す．

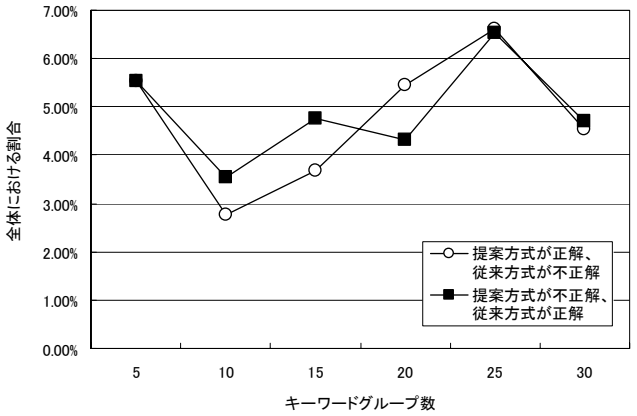


図 8:提案方式と従来方式との差異

提案方式と従来方式において予測内容が相反していた個所は平均 9.67%あった．そのうち提案方式が正解(従来方式が不正解)の割合は，平均 4.77%，提案方式が不正解(従来方式が正解)の割合は平均 4.90%であった．

(3)学習データ数による正答率の変化

次に「学習データ」として用いるコンテンツの数を変動させ，学習データ量と正答率の関係を確認した．「学習データ」として前半 5～70 コンテンツ（5 コンテンツ刻み）を用い，後半 30 コンテンツのコンテンツの選択結果を予測し，「正解データ」と比較して，正答率を算出した．図 9 にその結果を示す．各記号でプロットされている地点が提案方式の正答率である．実線は従来方式の正答率である．破線は全てを「削除」と予測した場合の正答率を示す．

図 9 より，プロットが右肩上がりの傾向を示してい

ることから，提案方式は学習データ数が多いと正答率が向上する傾向が見られた．しかし，使用したキーワードグループによっては，正答率が従来手法より上回る場合と下回る場合があり，最適なキーワードグループ数を一意に決定できないことが分かった．

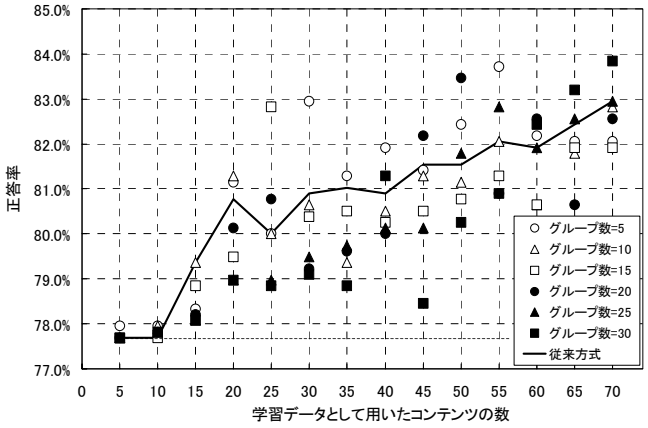


図 9:学習データ数による正答率の変化

3.3.2 再現率・精度・フォールアウト・R/F による評価

情報検索システムの性能評価の指標として用いられる再現率・精度等の指標を用いて評価を行った[10].

ここでは，選択実験結果の全 26 ユーザにおける前半 50 コンテンツを「学習データ」，後半 50 コンテンツを「正解データ」として，提案方式，従来方式を用いて全 26 ユーザの後半 50 コンテンツ（計 1300 個）の選択内容を予測した．

ここで，1300 個の予測のうち，正解データと予測内容が「保存」で一致した数を w，正解データと予測内容が「削除」で一致した数を z とする．また，正解データ中の「保存」の数を(w+x)，正解データ中の「削除」の数を(y+z)，予測内容が「保存」となった数を(w+y)，予測内容が「削除」となった数を(x+z) とする．なお，実験結果より，(w+x)=316，(y+z)=984，(w+x+y+z)=1300 はあらかじめ得られている．

表 5 に w,x,y,z の交差行列を示す．式(8)に再現率，精度，フォールアウト，R/F 値の算出式を表す．再現率 R，精度 P，フォールアウト F，R/F 値などは表 5 から算出できる．

表 5:交差行列

		予測内容	
		「保存」の数	「削除」の数
正解データ	「保存」の数	w	x
	「削除」の数	y	z

$$\left.\begin{array}{ll} \text{再現率} R=\dfrac{w}{w+x} & \text{フォールアウト} F=\dfrac{y}{y+z} \\ \text{精度} P=\dfrac{w}{w+y} & R/F\text{値}=\dfrac{\text{再現率} R}{\text{フォールアウト} F} \end{array}\right\} \quad (8)$$

式(8)において，再現率は，どれだけ漏れなく「保存」コンテンツを予測できたかを表している．精度は「保存」と予測されたコンテンツが，どれだけ正解データと一致していたかを表している．再現率と精度は大きいほど性能が良い．フォールアウトは，再現率と双対な尺度で，どの程度，正解データでは「削除」であるものを間違って「保存」と予測したかを表している．従って，フォールアウトは値が小さいほど性能が良いといえる．さらに，R/F 値はランダムな文書選択との比率を表し，R/F 値が 1 より小さい場合はランダムに予測した方が良いといえる．表 6 に再現率，精度，フォールアウト，R/F 値による評価結果を示す．この結果は，それぞれ正答率が最大となった時点における値である．

表 6:評価結果（正答率が最大における値）

	キーワードグループ数	再現率 R	精度 P	フォールアウト F	R/F 値
提案方式	5	0.320	0.743	0.036	8.986
	10	0.402	0.641	0.072	5.570
	15	0.329	0.667	0.053	6.228
	20	0.418	0.721	0.052	8.060
	25	0.329	0.738	0.038	8.753
	30	0.475	0.636	0.087	5.431
従来方式		0.491	0.635	0.090	5.423

この結果より，提案方式の再現率は従来方式よりもやや低い傾向がみられ，提案方式の精度は従来方式よりもやや高い傾向がみられた．再現率と精度はトレードオフの関係であり，提案方式は従来方式と比較して精度が高い半面，再現率が低い傾向がみられた．

R/F 値は，提案方式の平均が 7.17 で従来方式が 5.42 であり，提案方式が上回っており，さらに 1 よりも大きい値であった．予測内容はランダムに予測した場合よりも 7 倍程度正確であると言える．

図 10 に提案方式と従来方式の再現率－精度グラフを示す．このグラフは「保存」か「削除」かを判別するためのしきい値を変化させることで，それぞれを予測するコンテンツの数を変化させて得た．再現率が 1 に近づくにつれ，「保存」と予測するコンテンツの数が多く，逆に 0 に近づくにつれ「保存」と予測するコンテンツの数が少なくなる．なお，用いたキーワードグループは 20 個生成したものである．

図 10 から，提案方式は従来方法と比較して，再現率が 0.05～0.5 の区間，つまり後半 50 個のコンテンツのうち、「保存」として予測するコンテンツの数を 1 ユーザあたり平均 1～12 個の範囲において，従来方式よりも精度が上回っていた．再現率を 0.05 以下に絞った場合は，推薦数が 1 個を下回るため，評価はできない．

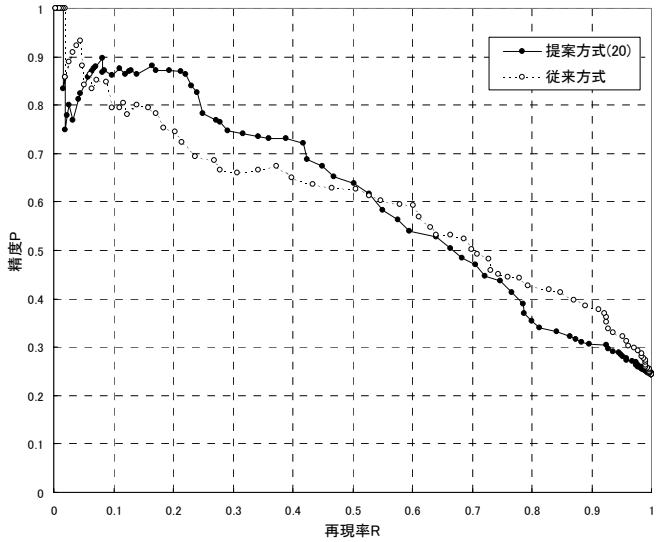


図 10:再現率－精度グラフ

3.3.3 まとめ

(1)提案方式と従来方式の正答率はほぼ同じであった.
(2)提案方式と従来方式との間で予測内容が全コンテンツ中の約 10%相違しており，Content-based フィルタリングでは予測できなかったコンテンツを約 5%選出できた.
(3)学習データ数の増加に伴って，提案方式と従来方式とではほぼ同様に正答率が上がっていった.
(4)R/F 値は,従来方式が 5.42 で,提案方式の平均が 7.17 であり，正確性の向上が確認できた.
(5)再現率を 0.05～0.5 にした場合（予測数を絞った場合)，従来方式よりも高い精度が得られた.

これらの結果から，Content-based フィルタリングによる単独ユーザの評価結果からでは類似性が判別できず予測できなかったコンテンツも他のユーザを反映することで予測できるようになった.

さらに，再現率が 0.05～0.5 の範囲（予測数を絞った場合）において高い精度が得られたことは，実際にユーザにコンテンツを推薦することを考えると，ユーザは大量のコンテンツを推薦することを望まないので,提案方式は有益であると言える.

4. おわりに

本研究では，複数ユーザのコンテンツ利用履歴から,キーワードグループを生成し，そのキーワードグループを介してユーザ・コンテンツ間類似度を算出する RUI-Filtering 方式を提案した．提案方式は，Content-based フィルタリングでありながら協調フィルタリングの効果を有することが確認でき，予測数を絞るといった条件によっては従来よりも高い精度を得ることができた.

今後，最適なメタ情報の量，キーワードグループの数について検討し，更なる評価によって提案方式の有効性や本方式が有効なサービスの検討，推薦されたコンテンツの意外性についての評価を行ってブラッシュアップを図る.

謝辞 日頃ご指導いただく NTT サイバーソリューション研究所一之瀬進所長，外村佳伸プロジェクトマネージャーに感謝いたします．また評価実験にご協力いただいた NTT サイバーソリューション研究所インテリジェントメディアプロジェクトの皆様に感謝いたします. またご指導を頂いた NTT サイバーソリューション研究所コンテンツ流通プロジェクトの谷口展郎様に感謝いたします.

参考文献

[1] F. Pachet, P. Roy, and D. Cazaly, "A Combinatorial Approach to Content-based Music Selection," IEEE Multimedia, vol. 7, pp. 44–51, January-March 2000.

[2] Goldberg, D., Nichols, D., Oki, B. M., and Terry, D., "Using Collaborative Filtering to Weave an Information Tapestry, " Communications of the ACM. December. 1992.

[3] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, "GroupLens: Applying Collaborative Filtering to Usenet News," Communicatons of the ACM. Vol. 40, No. 3, pp. 76-87, March 1997.

[4] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl, "GroupLens: An open architecture for collaborative filtering of netnews," In In Proceedings of the 1994 Computer Supported Cooperative Work Conference, pp. 175-186, New York, 1994. ACM.

[5] Marko Balabanovic and Yoav Shoham, "Fab: Content-based, collaborative recommendation, " Communications of the ACM, Vol. 40, No. 3, pp.66-72, 1997.

[6] 浅川智文，鎌原淳三，下條真司，宮原秀夫，川口知昭，山口陽一，“ユーザの嗜好の部分的な類似性を利用した情報推薦手法の提案,” 電子情報通信学会データ工学ワークショップ(DEWS2001), 5B-5，March.2001.

[7] 岡本道也，山下剛士，鎌原淳三，下條真司，宮原秀夫,“動的なカテゴリ定義を利用した個人化サービスの実現,” 電子情報通信学会データ工学ワークショップ(DEWS2002)，B3-6，March.2002.

[8] Badrul M. Sarwar, George Karypis, Joseph A. Konstan, John Reidl, "Item-Based Collaborative Filtering Recommendation Algorithms," 10th Int'l World Wide Web Conference, ACM Press, 2001, pp. 285-295.

[9] 土井俊介，塩原寿子，石黒正典,“通信放送融合型コミュニケーションサービスのためのユーザ選出手法,”電子情報通信学会データ工学ワークショップ(DEWS2002)，B3-5，March.2002.

[10] 徳永健伸, 情報検索と言語処理, 東京大学出版会, 1999.

[11] 形態素解析システム「茶筌」, http://chasen.aist-nara.ac.jp/