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

佐々木　祥†　　宮田　高道†　　稲積　泰宏††　　小林　亜樹†††　　酒井　善則†

† 東京工業大学 大学院理工学研究科　〒 152–8552 目黒区大岡山 2–12–1
†† 神奈川大学 工学部　〒 221-8686 横浜市神奈川区六角橋 3–27–1
††† メディア教育開発センター　〒 261–0014 千葉市美浜区若葉 2–12
E-mail: †{brnw,miyata,ys}@net.ss.titech.ac.jp, ††inazumi@cs.ee.kanagawa-u.ac.jp, †††aki@nime.ac.jp

あらまし　筆者らはソーシャルブックマークサービスにおけるタグ（自由記述のキーワード）を表象とする web コンテンツ群（以下，コンテンツクラスタ）の類似性に基づいた web コンテンツ推薦システムを提案した．この提案システムでは，タグの名称を用いないことで，ロングテールを構成する少数意見のタグ情報を有効に活用することに成功したが，一部の事例において推薦元のデータベース内で適合率の高い類似するコンテンツクラスタが不足していることに起因する推薦精度の低下が観測された．この原因について，他より広い概念に基づくコンテンツクラスタでは，それより狭いコンテンツクラスタから推薦を受けるため recall が低下し，逆に，狭い概念に基づくコンテンツクラスタでは，それより広いコンテンツクラスタから推薦を受けるため precision が低下する，などの分析を行った．そこで，この問題に対処するため，複数のコンテンツクラスタ間における論理演算によるクラスタの合成処理を導入することで，推薦に不足しているコンテンツクラスタの生成を試みる．本稿における事例調査の検証の結果，推薦精度の向上が期待できることを確認した．

キーワード　web コンテンツ推薦，ソーシャルブックマーク，コンテンツクラスタ

Content Recommendation System by Using Logical Operation of Contents Cluster

Akira SASAKI†, Takamichi MIYATA†, Yasuhiro INAZUMI††,

Aki KOBAYASHI†††, and Yoshinori SAKAI†

† Tokyo Institute of Technology　2–12–1 O-okayama Meguro-ku Tokyo, Japan, 152-8552
†† Kanagawa University　3–27–1 Rokkakubashi Kanagawa-ku Yokohama Kanagawa, Japan, 221-8686
††† National Institute of Multimedia Education　2–12 Wakaba Mihama-ku Chiba, Japan, 261–0014
E-mail: †{brnw,miyata,ys}@net.ss.titech.ac.jp, ††inazumi@cs.ee.kanagawa-u.ac.jp, †††aki@nime.ac.jp

Abstract　We proposed web recommendation system based on similarities among contents cluster, which is a group of web contents tagged with same keyword by user on social bookmark. this proposed method can use long-tail information of minor tag efficiently. However, in some experiments we found low recall or precision because of lack of suitable contents cluster marked high similarity. We revealed that recommendation from contents clusters based on narrower conception lead to low recall. We also revealed that recommendation from contents clusters based on wider conception lead to low precision. In order to solve this problem, we introduce logical operation of contents cluster to cover shortfall of them. In this paper we revealed those efficiency in some research cases.

Key words　Web Content Recommendation, Social Bookmark, Contents Cluster

1. はじめに

近年のインターネットでは爆発的な量の web コンテンツが短期間に生成されており，その中から必要とする情報を発見するために，ユーザは多大な労力を強いられている．この問題を解決するために，これまで数多くの web コンテンツ推薦システムが提案されている [1] [2] [3] [4]．これらの多くは，アイテムの共起性に基づいた推薦手法である協調フィルタリング [5] [6] [7] を技術的な核としている．協調フィルタリングの基本的なアルゴリズムは以下の通りである．

（1） ユーザによるアイテムの評価（購入，閲覧，採点など）を収集

（2） アイテムの評価傾向に基づくユーザ間の類似度を算出

（3） 類似度の高いユーザが高評価を与えたアイテムを推薦

協調フィルタリングは，アイテムの内容に依存することなく，ユーザの嗜好に基づいた推薦を可能とする．しかしながら現在，web 上には膨大な数の web コンテンツが存在しているため，これらを推薦対象にするためには大量のユーザの評価を収集する必要がある．

そこで著者らは文献 [8] において，近年流行しつつあるソーシャルブックマーク（以下，SBM）サービス [9] [10] を利用した web コンテンツ推薦手法を提案した．SBM とは，従来ローカルで行っていたウェブページのお気に入り登録であるブックマークを WWW 上で行うサービスの総称であり，このサービスの最大の特徴として，web コンテンツにタグと呼ばれるキーワードを付与できることが挙げられる．本手法ではユーザが web コンテンツに付与するタグの情報を，タグの名称としてではなく，タグ付与行動によって関連付けられた web コンテンツ集合（以下，コンテンツクラスタ）として利用することで，ロングテールを構成する少数意見のタグ情報を有効に活用することに成功した．しかしながら，一部の事例において推薦元のデータベース内で適合率の高い類似するコンテンツクラスタが不足していることに起因する推薦精度の低下が観測された．

そこで本稿では，文献 [8] で提案した web コンテンツ推薦手法の精度改善の一検討として，複数のコンテンツクラスタ間において論理演算によるコンテンツクラスタの合成を行うことで，推薦に不足しているコンテンツクラスタの生成を試みる．

本稿の全体の構成を以下に示す．まず 2 章で我々が提案している SBM を用いた web コンテンツ推薦手法の概要について説明する．次に 3 章で，論理演算によるコンテンツクラスタの生成について説明し，4 章において検証実験およびその結果を示す．最後に 5 章で本稿をまとめる．

2. SBM を用いた web コンテンツ推薦手法の概要

前述のとおり，著者らは SBM においてユーザが web コンテンツに付与するタグ（自由記述のキーワード）に着目した web コンテンツ推薦手法を提案している [8]．この手法において想定する推薦システムは，推薦を受けたい SBM ユーザが今まで付与してきたタグの中から興味のあるタグ名を入力することで，該当するコンテンツクラスタがクエリとして生成され，そのクエリと適合した web コンテンツが推薦されるというものである．以下では推薦手法の概要を説明する．

2.1 SBM のモデル化

あるユーザのブックマーク行動によって，全ての SBM 内に登録された web コンテンツは「ブックマークされている / ブックマークされていない」に分類される．また，ユーザによるタグの付与は，ブックマークしている web コンテンツに対して，タグを表象とする集合へ「帰属させる / 帰属させない」という二者択一を再度行っていることとみなせる．つまり，ある

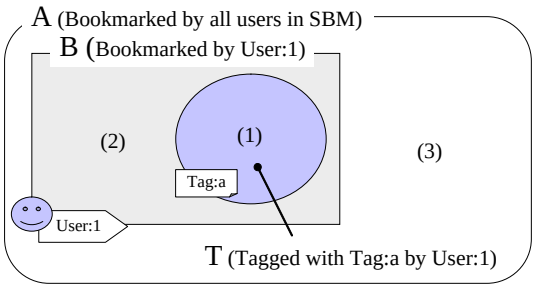

図 1　SBM におけるコンテンツ集合の関係

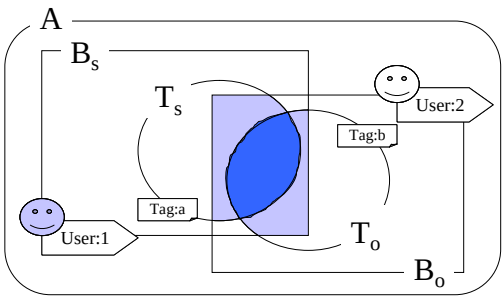

図 2　コンテンツクラスタの比較とコンテンツ推薦

特定のユーザ User:1 の，ある特定のタグ Tag:a に注目すると，SBM に登録されている全コンテンツ集合 A 内の web コンテンツは以下のいずれかに属している（図 1）．

（1） User:1 にブックマークされており，かつ，Tag:a が付与されている ($B \cap T$)

（2） User:1 にブックマークされているが，Tag:a は付与されていない ($B \cap \overline{T}$)

（3） User:1 にブックマークされていない ($\overline{B}$)

ユーザ User:1 によってタグ Tag:a を表象として結び付けられる web コンテンツ群を，User:1 の Tag:a による**コンテンツクラスタ**と呼ぶこととする．

本手法では，コンテンツクラスタの帰属関係が明示的に示されている集合 (1) および (2) の情報を元に，(3) にあたる未ブックマークの web コンテンツが当該コンテンツクラスタに帰属する可能性を提案アルゴリズムを用いて推薦度として算出し，推薦度の高さに基づいて web コンテンツの推薦を行う．

以下では，(1) に含まれる web コンテンツを**コンテンツクラスタに属する web コンテンツ**と呼び，(2) に含まれる web コンテンツを**コンテンツクラスタに属さない web コンテンツ**と呼ぶことにする．また，(3) に含まれる web コンテンツは，本提案における推薦対象となる web コンテンツである．

2.2 各ステップの説明

提案法のアルゴリズムは以下のステップで構成される．

（1） 二項分布に基づく尤度の算出

（2） 仮説検定に基づくコンテンツクラスタ間の類似度算出

（3） web コンテンツの推薦度算出

2.2.1 二項分布に基づく尤度の算出

図 2 は，二つのコンテンツクラスタ s(User:1, Tag:a)，o(User:2, Tag:b) における，web コンテンツの共起関係を示

したものである．ただし，s(User:1, Tag:a) は，ユーザ User:1 が付与したある特定のタグ Tag:a によるコンテンツクラスタを指すものとし，$B_{s(\text{User:1},*)}$ をユーザ User:1 によってブックマークされた web コンテンツ集合，$T_{s(\text{User:1,Tag:a})}$ をユーザ User:1 によってタグ Tag:a を付与された web コンテンツ集合とする．同様に，o(User:2,Tag:b) をユーザ User:1 とは異なるユーザ User:2 が付与したある特定のタグ Tag:b によるコンテンツクラスタとし，$B_{o(\text{User:2},*)}$ をユーザ User:2 によってブックマークされた web コンテンツ集合，$T_{o(\text{User:2,Tag:b})}$ はユーザ User:2 によってタグ Tag:b を付与された web コンテンツ集合とする．

以下では表記の省略のため，二つのコンテンツクラスタを s, o，二つのコンテンツクラスタに対応する集合を B_s，T_s，B_o，T_o の様に記述することとする．また，s はユーザによってクエリに指定されるコンテンツクラスタであるとして，**推薦先コンテンツクラスタ**と呼ぶこととする．o は推薦先コンテンツクラスタに対して web コンテンツを推薦するコンテンツクラスタであるとして，**推薦元コンテンツクラスタ**と呼ぶこととする．

ここではまず，コンテンツクラスタの確率的な帰属モデルを考える．本手法では，任意の web コンテンツのコンテンツクラスタへの帰属および非帰属は確率的に決定されると考える．二つのコンテンツクラスタ s, o において，任意の web コンテンツが少なくともいずれかのコンテンツクラスタに帰属する確率を $P(s \cap o)$ とし，両方のコンテンツクラスタに同時に帰属する確率を $P(s \cup o)$ とする．以上の確率を用いて，s，o の共起性の指標として一致度 p を次式で定義する．

$$p = \frac{P(s \cap o)}{P(s \cup o)} \tag{1}$$

次に，実際の SBM から測定されるデータについて，その結果が出現する確率，すなわち尤度を，先ほどの一致度 p に基づいて考える．s, o のいずれかに帰属する web コンテンツをランダムに $n(s,o)$ 個選択したとき，そのうち s, o のいずれにも同時に帰属する web コンテンツ数が $k(s,o)$ 個であったとする．ここで，任意の web コンテンツのコンテンツクラスタへの帰属関係は，他の web コンテンツと独立していると仮定すれば，その尤度は試行回数 $n(s,o)$，確率 p の二項分布に従う．よって，s, o に同時に帰属する web コンテンツの数が $k(s,o)$ 個となる尤度 L は一致度 p を用いて以下の式で与えられる．

$$L(n(s,o), k(s,o), p) = \begin{pmatrix} n(s,o) \\ k(s,o) \end{pmatrix} p^{k(s,o)} (1-p)^{n(s,o)-k(s,o)} \tag{2}$$

ただし，帰属関係を実際に比較することが出来るのは，s, o 間で同時にブックマークされている web コンテンツ群 $B_s \cap B_o$ のみである．そこで本手法では，$n(s,o)$, $k(s,o)$ を以下のように定義する．

$$n(s,o) = |(B_s \cap B_o) \cap (T_s \cup T_o)| \tag{3}$$

$$k(s,o) = |T_s \cap T_o| \tag{4}$$

つまり，二つのコンテンツクラスタにおいて，ブックマークの共通部分に属する web コンテンツ $B_s \cap B_o$ のうち，少なくとも一方のコンテンツクラスタに帰属する web コンテンツの個数 $|(B_s \cap B_o) \cap (T_s \cup T_o)|$ を試行回数 $n(s,o)$ とし，両方のコンテンツクラスタに帰属する web コンテンツの個数 $|T_s \cap T_o|$ を事象発生回数 $k(s,o)$ とする．すなわち，図 2 に示す二つの斜線部の和集合に含まれる web コンテンツの個数が $n(s,o)$ であり，積集合に含まれる web コンテンツの個数が $k(s,o)$ である．

2.2.2 仮説検定によるコンテンツクラスタ間の類似度算出

前節では SBM 上のすべての web コンテンツの帰属/非帰属が示されたコンテンツクラスタを仮定して尤度を定義した．しかしながら，実際には SBM にブックマークされている web コンテンツの情報しか用いることができない．そこでこのステップでは，二つのコンテンツクラスタ s, o において帰属/非帰属が示されている範囲での全コンテンツ，すなわち，$B_s \cup B_o$ における情報に基づき，式 (2) で得られた尤度を用いて一致度の仮説検定を行う．すなわち，二つのコンテンツクラスタ間の関係において，以下の 2 つの仮説を設定する．

- 一致度 p_1：共通の概念に基づく

共通の概念に基づいて構成されたコンテンツクラスタ同士では，web コンテンツが高確率で同時に帰属する．

- 一致度 p_0：異なる概念に基づく

異なる概念に基づいて構成されたコンテンツクラスタ同士では，web コンテンツが低確率で同時に帰属する．

ここで，二つのコンテンツクラスタ間の関係が,「共通の概念に基づく」・「異なる概念に基づく」のいずれかに分類されるとすると，どちらにより近いかが尤度比の大きさによって算出することが可能である．以下に，尤度比を算出する式を示す．

$$\begin{aligned} sim(s,o) &= \log \frac{L(n(s,o),k(s,o),p_1)}{L(n(s,o),k(s,o),p_0)} \\ &= k(s,o) \log \frac{p_1}{p_0} + (n(s,o) - k(s,o)) \log \frac{1-p_1}{1-p_0} \end{aligned} \tag{5}$$

式 (5) において $sim(s,o)$ は，値が大きいほど共通の概念に基づいていると検定される．よって本研究では，この対数尤度比 $sim(s,o)$ をコンテンツクラスタ間の**類似度**として利用する．この類似度に基づき，類似度が高いコンテンツクラスタから web コンテンツの推薦を行う．すなわち，T_o は T_s に対し，$\overline{T_s} \cap T_o$ に含まれる web コンテンツを類似度 $sim(s,o)$ にて推薦する．

ここで，本研究では p_0，p_1 の値として，多くの事例でよい結果が得られた数値 $p_0 = 0.1$，$p_1 = 0.6$ を採用した．このパラメータの最適値は，個々の事例毎に存在するが，事前の検証実験において，上記のパラメータの付近では推薦精度に大きな差異が無いことを確認している．

2.2.3 web コンテンツの推薦度算出

このステップでは，推薦先コンテンツクラスタ s に対する web コンテンツ c（推薦対象は図 2 における $\overline{B_s} \cap T_o$ である）の推薦度 $R(s,c)$ を，c を帰属している任意の推薦元コンテンツクラスタ $T_{o_i} (\forall i, c \in T_{o_i}, i = 1, 2, 3, \cdots, k)$ との類似度 $sim(s, o_i)$ の和によって定義する．

$$R(s,c) = \sum_{i=1}^{k} sim(s, o_i) \tag{6}$$

表 1　本実験処理後の web コンテンツの分類

	推薦集合 R	非推薦集合 r
正解集合 X	正解 RX	推薦漏れ rX
不正解集合 x	不正解 Rx	–

ただし，$sim(s,o) < 0$ となる値においては，前項の式 (5) の仮説検定において p_0 と判定されたとみなし，和に加えないこととする．

以上によって算出された値 $R(s,c)$ を，T_s に対する web コンテンツ c の**推薦度**と定義する．これに基づき，ランキングや閾値などの処理によってコンテンツの推薦を行う．

2.3 web コンテンツ推薦手法の検証実験

del.icio.us [9] から取得したデータを用い提案方式の実証実験を行った．同サービスは，世界中の多くのユーザに利用されている SBM の代表的な存在である．

この実証実験では，同サービスに登録しているユーザの情報を収集し，1,000 人分のブックマークデータを抽出した．このデータにおいて，総コンテンツ数は約 310,000 であり，延べタグ数は約 260,000 であった．すなわち，このデータセットにおける総コンテンツクラスタ数は約 260,000 である．本実験では，この 260,000 のコンテンツクラスタすべてを用いて推薦を行う．

2.3.1 実 験 方 法

本提案アルゴリズムは，ある推薦先コンテンツクラスタ s においてブックマークされていない web コンテンツの推薦を行うものである．しかしながら，ある web コンテンツが推薦されたとき，それが当該ユーザの所望するものであるかどうかを客観的に判定することは困難であるため，ここでは既に帰属関係が示されている web コンテンツ群，すなわち，当該ユーザのブックマーク B_s に含まれる web コンテンツ群に対して推薦度を算出することで，推薦結果と元のコンテンツクラスタとの web コンテンツの一致度によって推薦精度の検証を行う．

実験方法は以下のとおりである．

（1） 検証する推薦先コンテンツクラスタ s を選択，s に帰属している web コンテンツ（T_s）を正解集合 X，s に帰属していない web コンテンツ（$B_s \cap \overline{T_s}$）を不正解集合 x とする

（2） 他の全てのユーザによるコンテンツクラスタ $o_i (i = 1,2,3,\cdots)$ を対象に，s と o との間で類似度 $sim(s,o_i)$ を算出

（3） 算出した類似度をもとに，B_s に含まれる web コンテンツの推薦度を算出，上位数件を推薦集合 R，それ未満を非推薦集合 r とする

（4） 元の s と推薦結果を比較，recall および precision を算出

ただし，実験処理後の各 web コンテンツは，表 1 のいずれかのカテゴリに分けられるので，recall，precision は次の式で与えることとする．

$$\text{recall} = \frac{RX}{RX + rX} \tag{7}$$

$$\text{precision} = \frac{RX}{RX + Rx} \tag{8}$$

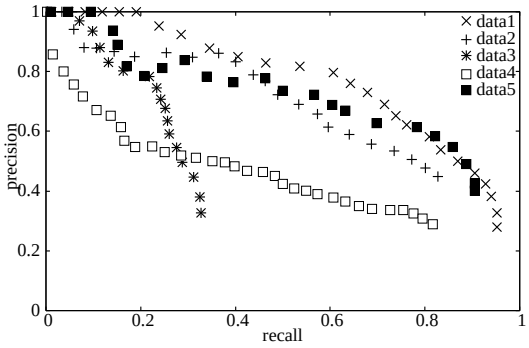

図 3　提案手法における recall-precision の関係 (推薦不可コンテンツ除去)

2.3.2 実 験 結 果

図 3 に，5 種類のコンテンツクラスタ（図中の data1〜data5）に対して前述の検証実験を行ったときの結果を示す．ただし，VC に含まれる web コンテンツが他の誰もブックマークしていない場合に，提案手法では推薦が不可能であるので，このことを考慮し，このような web コンテンツはそもそも推薦不可能な web コンテンツであるとして除去している．ここで検証に用いたコンテンツクラスタは，ランダムに選択した 5 名のユーザについて，それぞれのユーザが最も多くの web コンテンツに付与しているタグに対応するコンテンツクラスタである．それぞれのデータにおいて，ユーザおよびユーザがブックマーク登録した web コンテンツ数 ($|B|$)，ユーザが付与したタグの名称およびタグが付与された web コンテンツ数 ($|T|$)，その中から推薦不可能な web コンテンツを除いた web コンテンツ数 ($|T_{clear}|$) を表 2 に示す．

図 3 を見ると，おおむねの事例でよい結果が得られているが，data3 は recall が低く，data4 は precision が低くなっていることがわかる．

data3 の推薦先コンテンツクラスタは「art」であり，推薦元コンテンツクラスタは「webdesign」のタグ名に見るように，より限定的な概念でタグ付与していたと思われるものが多く見られた．data3 については，推薦先コンテンツクラスタに比べ，狭い概念に基づくと解される推薦元コンテンツクラスタがデータセット内に多く存在したため，推薦不足になり recall が低くなったものと考えられる．

また，data4 の推薦先コンテンツクラスタは「javascript」であり，推薦元コンテンツクラスタは「programming」のタグ名に見るように，より広範囲な概念でタグ付けしていたと考えられるものが多く見られた．data4 については，推薦先コンテン

表 2　実験結果におけるコンテンツクラスタの詳細

	data1	data2	data3	data4	data5
ユーザ	user1	user2	user3	user4	user5
$\|B\|$	2118	2459	17966	10079	4539
タグの名称	ajax	web	art	javascript	electronics
$\|T\|$	117	354	863	845	207
$\|T_{clear}\|$	84	272	449	427	106

ツクラスタに比べ，広い概念に基づく推薦元コンテンツクラスタがデータセット内に多く存在したため，過剰に推薦されて precision が低くなったものと考えられる．

3. 論理演算によるコンテンツクラスタの生成

前章まで述べた推薦手法は，検証実験によって高い精度で web コンテンツが推薦できることを確認した．しかしながら，一部の事例において推薦元のデータベース内に類似度の高いコンテンツクラスタが不足していることに起因する推薦精度の低下が観測された．

この原因について分析したところ，概念の広い推薦先コンテンツクラスタが概念の狭い推薦元コンテンツクラスタから推薦を受ける場合，推薦元コンテンツクラスタに属する web コンテンツのみ推薦を受けるため，recall が低下することがわかった．逆に，概念の狭い推薦元コンテンツクラスタが概念の広い推薦元コンテンツクラスタから推薦を受ける場合，推薦元コンテンツクラスタから過剰に推薦を受けるため，precision が低下したことがわかった．

そこで，この問題に対処するため，複数のコンテンツクラスタ間における論理演算によるクラスタの合成処理を導入する．これにより，データベース内に不足している推薦に適切なコンテンツクラスタの生成を試みる．

以下では，ユーザ num のタグ名 tagname によるコンテンツクラスタを，u_{num}[tagname] と表記する．

3.1 コンテンツクラスタの和演算

ここでは，コンテンツクラスタの和について検討する．これは,「Java」,「C」,「perl」などの狭い概念に基づくコンテンツクラスタをまとめることで,「プログラミング全般」という広い概念に基づくコンテンツクラスタの生成を行おうというものである．

u_1[a] に u_2[b] を和演算した結果，新たに生成されるコンテンツクラスタを u_1[a]+u_2[b] と表記することにする．ただし，ユーザ 1 とユーザ 2 は同一ユーザも可とする．

コンテンツクラスタ u_1[a]+u_2[b] において，コンテンツクラスタに属する web コンテンツ (図 1 における (1)) は，2 つのコンテンツクラスタに属する web コンテンツ集合の和集合 $T_{\mathrm{u}_1[a]} \cup T_{\mathrm{u}_2[b]}$ であり，コンテンツクラスタに属さない web コンテンツ (図 1 における (2)) は，2 ユーザによってブックマークされた web コンテンツ集合から (1) 部分を除いた集合 $(B_{\mathrm{u}_1[a]} \cup B_{\mathrm{u}_2[b]}) \cap (\overline{T_{\mathrm{u}_1[a]} \cup T_{\mathrm{u}_2[b]}})$ となる．

3.2 コンテンツクラスタの差演算

次に，コンテンツクラスタの差について検討する．これは,「プログラミング全般」という広い概念に基づくコンテンツクラスタから目的以外の具体的なプログラミング言語を取り除くことで,「Java」,「C」,「perl」などの狭い概念に基づくコンテンツクラスタの生成を行おうというものである．

u_1[a] から u_2[b] を差演算した結果，新たに生成されるコンテンツクラスタを u_1[a]−u_2[b] と表記することにする．ただし，ユーザ 1 とユーザ 2 は同一ユーザも可とする．

コンテンツクラスタ u_1[a]−u_2[b] において，コンテンツクラスタに属する web コンテンツ (図 1 における (1)) は，2 つのコンテンツクラスタに属する web コンテンツ集合の差集合 $T_{\mathrm{u}_1[a]} \cap \overline{T_{\mathrm{u}_2[b]}}$ であり，コンテンツクラスタに属さない web コンテンツ (図 1 における (2)) は，ユーザ u_1 によってブックマークされた web コンテンツ集合から (1) 部分を除いた集合 $B_{\mathrm{u}_1[a]} \cap (\overline{T_{\mathrm{u}_1[a]} \cap \overline{T_{\mathrm{u}_2[b]}}})$ となる．

4. コンテンツクラスタの論理演算の検証実験

コンテンツクラスタの論理演算によって生成されるコンテンツクラスタが，web コンテンツ推薦の精度および再現率の改善を達成できるかを検証するため，以下の実験を行う．

4.1 実験方法

2.3 に示した検証実験において，精度または再現率が低くなったコンテンツクラスタを推薦先コンテンツクラスタとして選択する．このとき，推薦元コンテンツクラスタのいくつかを選択し，論理演算によって合成コンテンツクラスタを生成し，推薦元コンテンツクラスタとの類似度に改善が見られるかを検証する．

4.2 実験結果

表 3 において，推薦先コンテンツクラスタ u_0[programming] と，二つの推薦元コンテンツクラスタ u_1[javascript],u_1[ajax] との類似度および二つのコンテンツクラスタを和演算したコンテンツクラスタ u_1[javascript]+u_1[ajax] との類似度を示す．このように，広い概念に基づいた推薦元コンテンツクラスタ u_0[programming] に対して，従来手法では u_1[ajax] の web コンテンツを推薦することができなかったが，論理演算によって合成されたコンテンツクラスタ u_1[javascript]+u_1[ajax] によって u_1[ajax] の web コンテンツも推薦することが可能となり，recall の改善が期待できると考えられる．

また，表 4 は，推薦先コンテンツクラスタ u_2[javascript] と，二つの推薦元コンテンツクラスタ u_3[programming],u_4[ruby] との類似度および二つのコンテンツクラスタを差演算したコンテンツクラスタ u_3[programming]−u_4[ruby] との類似度を示す．このように，狭い概念に基づいた推薦元コンテンツクラスタ u_2[javascript] に対して，従来手法では他のプログラミング言語である「ruby」について記述された web コンテンツまで含めて推薦が行われていたが，論理演算によって合成された u_3[programming]−u_4[ruby] では，過剰な推薦を抑えるとともに

表 3　和演算によるコンテンツクラスタの類似度変化比較

推薦元コンテンツクラスタ o	$n(s,o)$	$k(s,o)$	$\frac{k(s,o)}{n(s,o)}$	$sim(s,o)$
u_1[javascript]	57	22	0.39	+11.04
u_1[ajax]	59	14	0.24	−11.41
u_1[javascript]+u_1[ajax]	59	30	0.51	+30.24

表 4　差演算によるコンテンツクラスタの類似度変化比較

推薦元コンテンツクラスタ o	$n(s,o)$	$k(s,o)$	$\frac{k(s,o)}{n(s,o)}$	$sim(s,o)$
u_3[programming]	57	22	0.39	+11.04
u_4[ruby]	33	1	0.03	−24.16
u_3[programming]−u_4[ruby]	53	22	0.42	+14.28

に,「ruby」以外のプログラミング言語について記述された web コンテンツの推薦度を向上させるため，precision の改善が期待できると考えられる．

4.3 コンテンツクラスタの合成基準について

今回提案したコンテンツクラスタの合成は，無作為に選択したコンテンツクラスタ組に対して行ったとしても必ずしもよい結果を得られるわけではない．そこで，合成することによって推薦精度が向上するコンテンツクラスタ組の発見法についての検討が必要となる．

考えられる方法として，ある推薦先コンテンツクラスタに対し，何らかの基準（web コンテンツ数，web コンテンツの共起性など）に基づいて合成する推薦元コンテンツクラスタ組候補を選び出し，それらの間で実際に合成を行ってみる，という方法が挙げられる．すなわち，合成候補であるいくつかの推薦元コンテンツクラスタ間で合成を行い，実際に推薦先コンテンツクラスタとの類似度計算を行った結果，本実験の例に挙げられるような類似度の改善が見られた場合にのみ適用する．

しかしながらこの方法は，合成候補の推薦元コンテンツクラスタを効率的に発見する基準を十分検討する必要がある．また，コンテンツクラスタの合成による推薦速度の低下などが今後の課題となる．

5. おわりに

本稿では，文献 [8] で提案した web コンテンツ推薦手法の精度改善の一検討として，コンテンツクラスタの論理演算を導入し，新たなコンテンツクラスタの生成を行った．検証実験により，生成されたコンテンツクラスタが web コンテンツ推薦の精度を向上することが確認できた．今後の課題は，推薦精度の変化の検討や，論理演算によって推薦精度が改善されるコンテンツクラスタ組の発見手法の詳細な検討や，推薦における速度の向上の検討などが挙げられる．

文　献

[1] J. Li, O. Zaiane, “Combining Usage, Content, and Structure Data to Improve Web Site Recommendation,” Proceedings of Web KDD-2004 workshop on Web Mining and Web Usage, 2004.

[2] P. Kazienko, M. Kiewra, “Integration of Relational Databases and Web Site Content for Product and Page Recommendation,” International Database Engineering and Applications Symposium (IDEAS’04), 2004.

[3] H. Ishikawa, T. Nakajima, T. Mizuhara, S. Yokoyama, J. Nakayama, M. Ohta, K. Katayama, “An Intelligent Web Recommendation System: A Web Usage Mining Approach,” International Symposium on Methodologies for Intelligent Systems, pp.342-350, Jun. 2002.

[4] S. gunduz, M. T. ozsu, “A user interest model for web page navigation,” Proceedings of International Workshop on Data Mining for Actionable Knowledge, Apr. 2003.

[5] D. Goldberg, D. Nichols, B. M. Oki, D. Terry, “Using collaborative filtering to weave an information tapestry,” Communications of the ACM, Vol. 35 No. 12, pp. 61–70, 1992.

[6] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, “GroupLens: An Open Architecture for Collaborative Filtering of Netnews,” Proceedings of the 1994 Computer Supported Cooperative Work Conference, pp 175–186, 1994.

[7] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl. “Item-based collaborative filtering recommendation algorithms,” Proceedings of the 10th International World Wide Web Conference (WWW10), pp 285–295, 2001.

[8] 佐々木祥，宮田高道，稲積泰宏，小林亜樹，酒井善則，“Social 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，”Proceedings of DBWeb2006, pp 59–66，2006.

[9] del.icio.us，http://del.icio.us/

[10] はてなブックマーク，http://b.hatena.ne.jp/