情報処理学会論文誌　データベース　Vol. 2　No. 2　44–55 (June 2009)

プロファイル間類似度の推移関係に着目した推薦計算量削減

佐々木　祥[†1]　宮田高道[†1]　稲積泰宏[†2]
小林亜樹[†3]　酒井善則[†1]

アイテム推薦手法の多くで用いられている協調フィルタリングでは，規模の拡大にともない，プロファイル間の類似度計算やアイテムの推薦度計算の回数が増加し，推薦時間が長くなるという問題がある．そのため，推薦に有用となる類似度の高いプロファイルを選択することで，推薦精度を保ちつつ計算量を減らすことが必要となる．そこで本研究では，プロファイル間の類似度において推移関係があることに着目した．各プロファイルをノードとするプロファイル選択ネットワークを作成し，類似度に従う確率的探索を行うことによって，有用なプロファイルを発見する手法を提案するとともに，ソーシャルブックマークの実データを用いて本手法の有効性を示す．

Calculation Time Reduction in Item Recommendation System Based on Transitive Law of Similarities between Each Profiles

Akira Sasaki,[†1] Takamichi Miyata,[†1]
Yasuhiro Inazumi,[†2] Aki Kobayashi[†3]
and Yoshinori Sakai[†1]

Scalability is the biggest problem if one wants to implement item recommendation system in real world. Increasing of items and users leads to increse the calculation time and reduce the efficiency of recommendation system. Therefore, selection of effective recommender profiles which are similar to recommendee is required to reduce calculation time. In this paper, we focused on 'transitive law' of similarities between each profiles and replace the scalability problem into a node searching problem in a pseudo-distributed network. Experimental results, based on live data from real social bookmark service, shows that our proposed method have potential to reduce cal culation time drastically and select the effective profile from distributed network.

1. はじめに

近年，blog や web 日記を代表とするユーザ参加型サービスが広く普及したことにともない，WWW 上のコンテンツは増加傾向にあり，ユーザは所望するコンテンツを発見することが困難となってきている．このような状況を受け，コンテンツの発見を容易にするコンテンツ推薦システムが注目されており，これまで数多くの提案がされてきた．これらの多くは，大量の情報から有用な情報を選択する情報フィルタリングの一手法である協調フィルタリング[1)–3)]を技術的な核としている．

協調フィルタリングは，ユーザから収集した各アイテムの有用性（ユーザによるアイテムの 5 段階評価など）の情報列である嗜好情報（プロファイル）の類似性に基づいて未知のアイテムの有用性を推測する手法であり，アイテムの内容に依存しないため，多くのアイテム推薦システムに用いられている．ここでは，以下の手順を施すことを協調フィルタリングと呼ぶこととする．

(1)　注目するプロファイル p_s と他の全プロファイル p_{o_k}（$k = 1, 2, \ldots$）間の類似度を，アイテムの有用性の傾向の類似性から算出．

(2)　p_s が有用性を評価していないアイテム i の有用性を，i を評価済みのプロファイルとの類似度に基づき評価．

以上のような全プロファイルとの比較を行う推薦手法では，プロファイル数やアイテム数の増加にともない計算量が増加するため，そのスケーラビリティの低さが問題となる．これに対し，注目するプロファイルとその他の全プロファイル間で類似度を計算し，類似度が一定以上となるプロファイル集合（以下，**近傍プロファイル**）のみを用いて推薦しても精度が保たれることが指摘されている[4)]．しかしながら，厳密に近傍プロファイルを求めるには全プロファイル間の類似度計算を必要となってしまう．よって，なるべく類似度計算をせずに，近似的に近傍プロファイルを求めることが必要となる．

そこで本稿では，全プロファイル中から**推定近傍プロファイル**を選択し，その間でのみ類

†1 東京工業大学
Tokyo Institute of Technology
†2 富山大学
University of Toyama
†3 工学院大学
Kougakuin University

© 2009 Information Processing Society of Japan

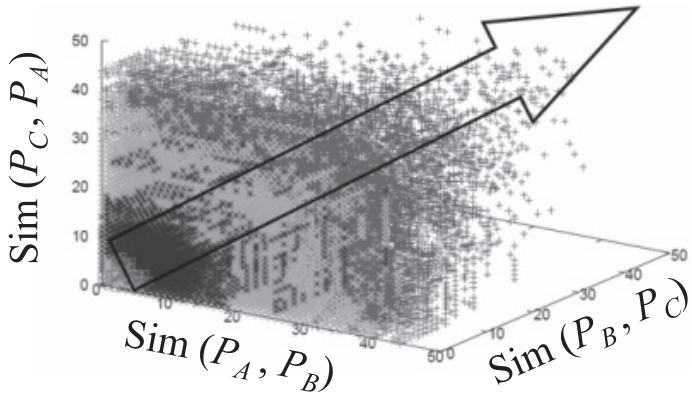

図 1　実データにおける類似度の 3 者関係
Fig. 1　Distribution of similarities among three profiles.

似度を計算することにより計算量を削減する手法を提案する．すなわち，少量の類似度計算に基づいて，注目するプロファイルとの類似度が高いと予測されるプロファイルの集合を選択することにより，検索精度を保ちつつ計算量を削減する手法を提案する．

ところで，実データを考察すると，あるプロファイル p_A と p_B 間の類似度 $\mathrm{sim}(p_A, p_B)$ が高く，また，p_B と p_C 間の類似度 $\mathrm{sim}(p_B, p_C)$ が高いとき，p_A と p_B 間の類似度 $\mathrm{sim}(p_C, p_A)$ もまた高いことが多いという，いわば「類似度の推移関係」が見られた．**図 1** は，実データから取得した類似度の分布であり，任意の 3 つのプロファイル p_A，p_B，p_C を取り出し，その 2 者間 (p_A, p_B)，(p_B, p_C)，(p_C, p_A) の類似度 $\mathrm{sim}(p_A, p_B)$，$\mathrm{sim}(p_B, p_C)$，$\mathrm{sim}(p_C, p_A)$ を，我々が以前行った既存研究 5) で提案した類似度計算法を用いて算出し，点 $(\mathrm{sim}(p_A, p_B), \mathrm{sim}(p_B, p_C), \mathrm{sim}(p_C, p_A))$ をプロットしたものである．

しかしながら，この図からただちに前述の推移関係の有無を調べることは難しい．そこで，全プロファイル間の類似度からランダムに 3 つの類似度 $\mathrm{sim}(p_A, p_B)$，$\mathrm{sim}(p_C, p_D)$，$\mathrm{sim}(p_E, p_F)$ を取り出し，それぞれを上記と同様の 3 者間の類似度の分布と “見なし” てプロットしたものを**図 2** に示す．この図では $\mathrm{sim}(p_A, p_B)$，$\mathrm{sim}(p_C, p_D)$，$\mathrm{sim}(p_E, p_F)$ は独立となっている．図 1 と図 2 を比較すると，図 2 では 3 つの類似度がすべて高くなる点はほとんど存在しないのに対して，図 1 ではそのような点が数多く見られることから，前述の 3 者間の推移関係の存在が確認できる．

本稿ではこの傾向を利用して，少量の類似度計算によって得られた，近傍プロファイルの

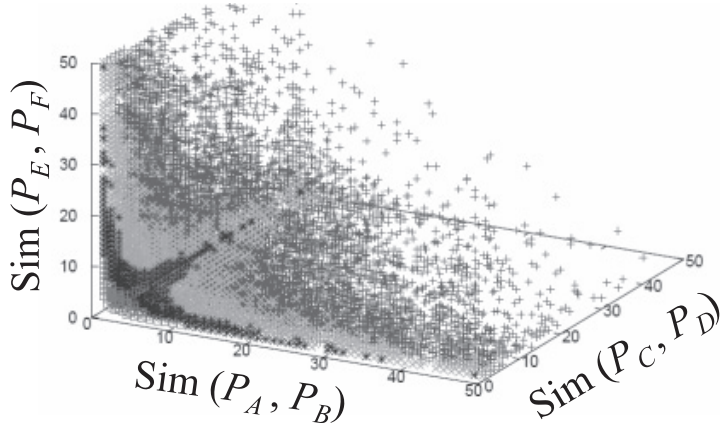

図 2　ランダムに設定した類似度の 3 者関係
Fig. 2　Distribution of similarities of randomly picked up three profile pairs.

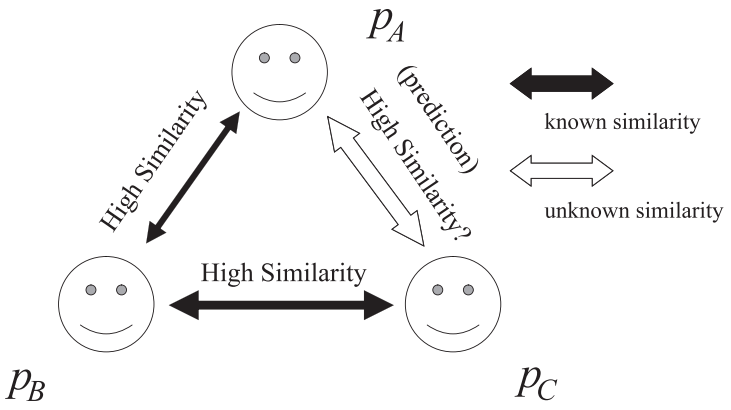

図 3　類似度の推移関係に基づく類似度予測
Fig. 3　Prediction based on transitive law of similarities.

部分集合から未知の近傍プロファイルを推移関係から予測し，推定近傍プロファイルとして選択する（**図 3**）．これにより，類似度計算の回数を削減しつつ，かつ，精度・再現率の高い推薦を実現する．

© 2009 Information Processing Society of Japan

表 1　関連研究
Table 1　Conventional studies.

計算方法	代表的手法	計算資源の利用	連続大量処理時	解法
逐次計算法	naive [2]	分散，削減なし	処理量に比例	完全
事前計算法	キャッシュ	時間的に分散	事前計算不可能	完全
並列計算法	pLSA 法[6]	空間的に分散	時間短縮不可能	完全
提案計算法		論理的に削減	時間短縮効果	近似

2. 関 連 研 究

本手法は計算量そのものの削減を目的とする手法である．本章では，提案手法と**表 1** に示したいくつかの関連研究との比較により，そのアプローチの違いを明らかにする．

前述のように，協調フィルタリングの実利用においては計算量の多さが主要な問題となる．各推薦処理に対して毎回類似度計算を行う単純な逐次計算法[2] ではこの問題は特に顕著となり，サービス面ではユーザの待ち時間の長さとして，運用面では計算機台数の増加，使用電力量の増加などの運用コストの増加として現れる．そのため，サービス面での問題に特化した解決法として，計算結果のキャッシングに代表される事前計算法と，pLSA 法[6] に代表される多数の小問題に分割して並列計算を行う手法が考えられてきた．

事前計算法では，将来的に必要となるであろう計算結果を事前に計算しておく方式であり，計算資源を時間軸方向で分散させて利用していることになる．同一の処理内容が多数回繰り返されるような状況で大きな力を発揮し，二度目以降の計算量を劇的に削減する効果が見込める．たとえば，文書検索における索引語による転置インデックスはこの代表例といえ，固定された検索対象文書集合，繰り返し用いられる（全数に対して相対的に）少数の索引語，という場面において効果をもたらしてきた．

一方で，協調フィルタリングにおいては，クエリ，あるいは演算対象となる履歴情報に変化があった場合には単純な事前計算では対処できない．クエリは，索引語のように固定的な表現が見込まれるわけではなく，利用者の嗜好に対処するために非常に高次元の情報となると考えられ，その場合の数は莫大であり，ストレージ容量の観点からも事前計算には向かない．また，履歴情報も利用者による一利用ごとに変化するものであり，変化の反映を集約するなどしても，従来の文書検索ほど固定化された状況ではないと考えられる．このような状況では，2 度目以降の計算量の削減効果は小さく見積もらざるをえず，時間軸方向での単なる処理負荷の移動にすぎなくなる．そればかりか，つねに変化する多様な計算処理に対処するための，将来を想定した事前計算はほとんど不可能であり，仮にすべての可能性を網羅しようとしたときには，逐次計算法とほぼ変わらない計算量となってしまう．このように，事前計算法は協調フィルタリングにおける計算量削減策とはなりえず，また，計算時間削減の面からも効果は薄いといわざるをえない．

並列計算法は，計算処理を分割して複数の計算機上で同時に実行することでサービスにおける待ち時間を短縮しようとする．たとえば文献 8) は，pLSA 手法を取り入れた実装例であり，EM アルゴリズム[7] の並列実行によって推薦計算処理を複数台の計算機に分散させる．これは，いわば空間方向に計算処理を分散する手法であるといえる．しかし，計算量自体の削減とはなっておらず，連続大量処理を想定した場合に，並列度に応じたサービス時間の削減効果を得るためには，逐次計算法において待ちが発生しないのと同等の計算資源を投入する必要がある．したがって，規模拡大に対してサービスを運用するためのコストは増加の一途をたどることになってしまう．

これに対して，提案方式は近似的な解法を示すものであり，本質的な計算量の削減をもたらす手法である．そのため，単にサービス時間の短縮が可能なばかりでなく，運営面から見ても，全体の処理負荷に対して必要となるコストの削減を実現する．

3. 提 案 手 法

本手法では事前情報として，データベース中の**いくつかの**プロファイル間において，すでに類似度が求められていることを前提とする．これは，各プロファイルにおいて，過去の推薦の履歴（**キャッシュ**）を利用できる状況と同様であると考えることができる．

本稿の目的は，適切なプロファイル選択を行うことにより推薦に必要な計算量を削減することである．そこで本手法では，事前情報に基づいて仮想的なネットワークを作り，このネットワーク上をクエリを hop-by-hop で伝搬させることによってプロファイル選択を試みる．以下では，このネットワークを**プロファイル選択ネットワーク**と呼ぶこととする．

プロファイル選択ネットワークにおいて，ノードは各プロファイルと対応しており，リンクは当該プロファイル間の類似度が既知であること，すなわち，キャッシュを保持していることに対応する．プロファイル間の類似度の定義は，協調フィルタリングの手法によって異なるため，つねに対称性が保証されているとは限らない．そこで，このリンクは有向であるとし，リンク元が類似度の情報を持つプロファイルに対応するノード，リンク先がリンク元との類似度を算出済みのプロファイルに対応するノードであるとする．

今回想定する推薦システムでは計算量を制御することを目的としているため，計算量との

© 2009 Information Processing Society of Japan

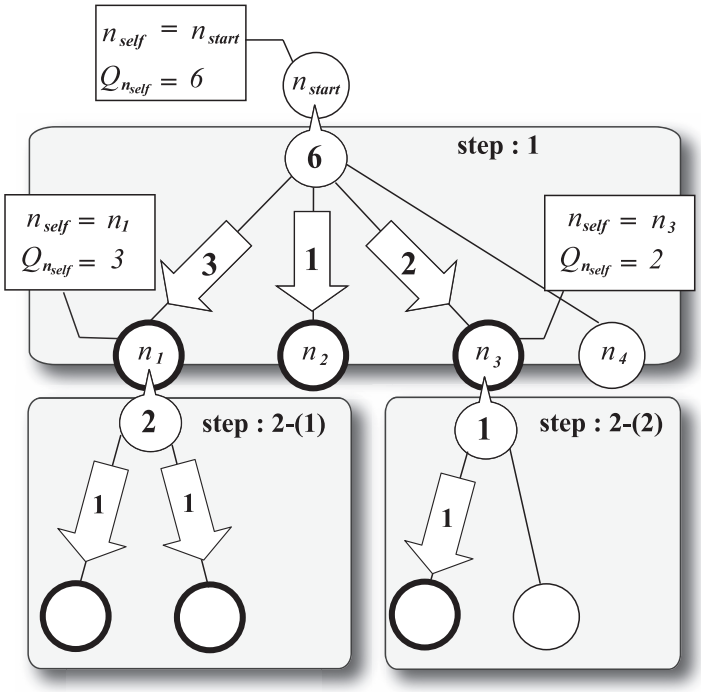

図 4　ノード探索
Fig. 4　An example of nodes searching.

関わりが大きいと考えられる，推薦時に比較するプロファイル数をシステムへの入力とする．つまり，このネットワーク上では，選択するプロファイルの個数が推薦時のクエリとして入力される．以下，これを**比較数** Q とする．

次に，プロファイルを再帰的に選択するアルゴリズム（以下，**探索**）を，**図 4** を用いて説明する．ここでは，ノード n_{start} の比較数 6 の探索から始まる状況を考える．step: 1 では，n_{start} と直接リンクしている n_1，n_2，n_3，n_4 に対して，クエリを比較数の個数だけ分配する．

- 探索 $(n_{start}, 6)$

この結果，$(n_1, n_2, n_3, n_4) = (3, 1, 2, 0)$ のようにクエリが分配されたとする．このとき，クエリを受け取ったノード n_1，n_2，n_3 は，受け取った比較数の探索を再帰的に行う．

```
クエリ発生ノード n_start;
初期の比較数 Q_start;
推定近傍プロファイル nodelist;
Search(n_start , Q_start);
 Search(n_self, Q_{n_self})
   自ノード n_self;
   比較数 Q_{n_self};
   リンク先のノード n_1, n_2, ···, n_i;

   if n_self が nodelist に無い then
     add n_self → nodelist;
     Q_{n_self} = Q_{n_self} − 1;
   while(Q_self > 0)
     n_x =SelectNode(n_self);
     Q_{n_x} = Q_{n_x} + 1;
     Q_{n_self} = Q_{n_self} − 1;
   for all 1 ≤ x ≤ i do
     Search(n_x, Q_{n_x});
```

図 5　プロファイル選択のアルゴリズム
Fig. 5　Algorithm of profiles selection.

- 探索 $(n_1, 3)$
- 探索 $(n_2, 1)$
- 探索 $(n_3, 2)$

以下の探索では，まず自己を推薦に利用するプロファイルを list に返し（同時に比較数を 1 減らす），比較数 2 以上のクエリを受け取った n_1，n_3 は，孫ノードに残りの比較数個分のクエリを分配する（step: 2-(1)，2-(2)）．以上のような再帰的な処理により，list に記された太い黒丸のノードと対応するプロファイルを用いて推薦を行う．以上のアルゴリズムを**図 5** に記す．

本稿では，上記探索において，**図 6** に示す方式を提案する．これはすなわち，リンク先のノードとの類似度の比率に応じた確率的な探索である．以下，これを**提案探索方式**と呼ぶ．これによって，注目するプロファイルと十分類似度は高いが，未知である有用なプロファイル集合，すなわち推定近傍プロファイルを，仲介するプロファイルを通して発見できるようになると考えられる．

一方，**図 7** に示すような単純なランダムによる探索の方式が考えられる．これはすなわち，どのリンク先のノードも等確率で選択されることとなる．以下，これを**ランダム探索方**

© 2009 Information Processing Society of Japan

提案探索方式
SelectNode(n_{self})
　自ノード n_{self};
　n_{self} のリンク先のノード $n_1, n_2, \cdots, n_i$;
　リンク先のノードとの類似度
　　$\mathrm{sim}_{n_1}, \mathrm{sim}_{n_2}, \cdots, \mathrm{sim}_{n_i}$;
　ランダム変数 $r(0 \leq r < 1)$;

　for all $1 \leq x \leq i$ **do**
　　$c = \sum_{j=1}^{x} \mathrm{sim}_{n_j} / \sum_{j=1}^{i} \mathrm{sim}_{n_j}$;
　　if $(c > r)$ **then**
　　　return (n_x);
　　end;

図 6　提案探索方式
Fig. 6　Proposed method of the node searching.

ランダム探索方式
SelectNode(n_{self})
　自ノード n_{self};
　n_{self} のリンク先のノード $n_1, n_2, \cdots, n_i$;
　ランダム変数 $r(0 \leq r < 1)$;

　for all $1 \leq x \leq i$ **do**
　　$c = x/i$;
　　if $(c > r)$ **then**
　　　return (n_x);
　　end;

図 7　ランダム探索方式
Fig. 7　Random node searching.

式と呼ぶ．以下の実験では，このランダム方式に対して，提案方式のほうが推薦精度において有効であることを示す．

4. 実　　験

本実験ではノード選択方式において，提案探索方式およびランダム探索方式を用いた検証を行い，提案探索方式の有用性を示す．

4.1 実 験 方 法

本稿ではソーシャルブックマーク（Social Bookmark，以下 SBM[9),10)]）サービスの delicious から，2008 年 6 月に収集したブックマークデータを検証実験に利用する．まず，ブックマーク数が 1000 件以上のユーザ 300 人について，それらのユーザがブックマークしたコンテンツと，各ユーザがコンテンツに付与したタグのデータを取得する．本検証実験では協調フィルタリングの手法として，我々が以前文献 5) において提案した SBM のタグ情報を利用した手法を用いるが，この手法では，各ユーザが付与したタグごとに注目し，それぞれを個別の**プロファイル**と見なす．すなわち，通常の協調フィルタリングとは異なり 1 人のユーザが複数のプロファイルを持ちうる．そこで，前述のユーザ 300 人について，ユーザごとに最も多く付与したタグを選び，当該タグが付与されたコンテンツの集合を，以下の検証実験でプロファイルとして用いる．この処理により，ユーザ数と同数の 300 個のプロファイルを得た．

4.2 タグ情報を利用した協調フィルタリングの概要

ここでは，我々の既存研究である，タグ情報を利用した協調フィルタリングについて概要を説明する．SBM とは，WWW 上でブックマークを管理および共有できるサービスであり，その利便性により現在急速に利用者数を増やしている．このサービスの最大の特徴として，コンテンツ（URL）に「blog」，「面白い」などといった自由記述によるキーワードであるタグを付与して記録できることがあげられる．

SBM では，あるユーザのブックマーク行動によって，すべての SBM 内に登録されたコンテンツは「ブックマークされている/ブックマークされていない」のどちらかに分類される．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見なせる．つまり，ある特定のユーザ u_0 の，ある特定のタグ t_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 内の任意のコンテンツは，以下のいずれかの集合に属している（**図 8**）．

(1)　ユーザ u_0 にブックマークされており，かつ，タグ t_a が付与されている（$B_s \cap T_s$）．
(2)　ユーザ u_0 にブックマークされているが，タグ t_a は付与されていない（$B_s \cap \overline{T_s}$）．
(3)　ユーザ u_0 にブックマークされていない（$\overline{B_s}$）．

ただし，B_s とは，ユーザ u_0 によってブックマークされたコンテンツ集合であり，T_s とは，ユーザ u_0 によってタグ t_a を付与されたコンテンツ集合を指す．以下本稿では，ユーザ u_0 によってタグ t_a を表象として結び付けられるコンテンツ集合を，u_0 の t_a による**コンテンツクラスタ**と呼ぶこととする．

ここで，前述のコンテンツクラスタと，他のユーザ u_i の t_b によるコンテンツクラスタを

© 2009 Information Processing Society of Japan

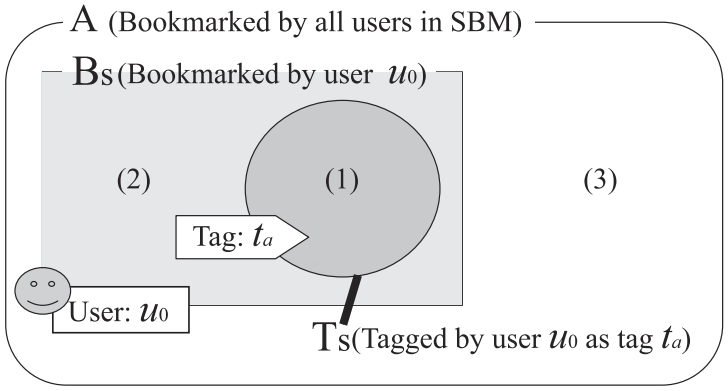

図 8　SBM におけるコンテンツ集合の関係
Fig. 8　SBM modeling about relationship among items, users, and tags.

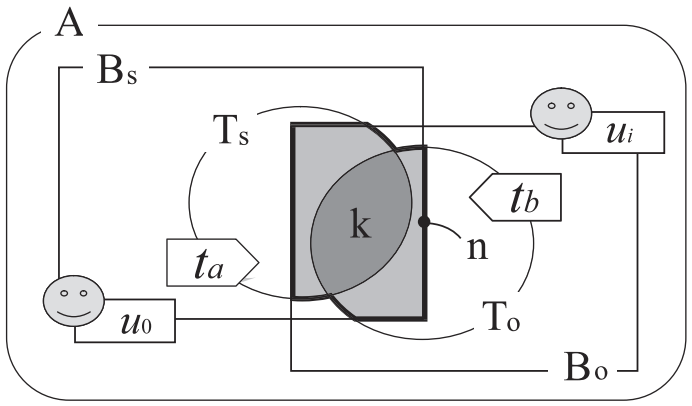

図 9　コンテンツクラスタの比較とコンテンツ推薦
Fig. 9　Item recommendation by comparison of item clusters.

比べたとき，2 つのコンテンツクラスタの関係は一般に**図 9** のようになる．ただし，B_o とは，ユーザ u_i によってブックマークされたコンテンツ集合であり，T_o とは，ユーザ u_i によってタグ t_b を付与されたコンテンツ集合を指す．このとき，それぞれの共通部分の個数 n, k を以下のとおりに定義する．

$$n(T_s, T_o) = |(B_s \cap B_o) \cap (T_s \cup T_o)| \tag{1}$$

$$k(T_s, T_o) = |T_s \cap T_o| \tag{2}$$

既提案法では二項分布の考えに基づいて，n 個をサンプリング数とし，そのうち k 個が成功する尤度 $L(n(T_s, T_o), k(T_s, T_o), p)$ を以下のように算出した．

$$\begin{aligned} &L(n(T_s, T_o), k(T_s, T_o), p) = \\ &{}_{n(T_s,T_o)}\mathrm{C}_{k(T_s,T_o)} \\ &\times p^{k(T_s,T_o)} (1-p)^{n(T_s,T_o)-k(T_s,T_o)} \end{aligned} \tag{3}$$

ここで，p を 2 つのコンテンツクラスタの一致度と呼び，任意のコンテンツに対して 2 人のユーザが同じコンテンツクラスタに帰属させる確率であると考える．そして，2 つの一致度として，概念が似ている p_1 および，似ていない p_0 を仮定し，p がいずれに近いか尤度比検定によって評価し，その値の高さによって 2 つのコンテンツクラスタの類似度 $\mathrm{sim}(T_s, T_o)$ を算出した．

$$\begin{aligned} &\mathrm{sim}(T_s, T_o) \\ &= \log \tfrac{L(n(T_s,T_o),k(T_s,T_o),p_1)}{L(n(T_s,T_o),k(T_s,T_o),p_0)} \\ &= k(T_s, T_o) \log \tfrac{p_1}{p_0} \\ &\quad +(n(T_s, T_o) - k(T_s, T_o)) \log \tfrac{1-p_1}{1-p_0} \end{aligned} \tag{4}$$

注目するコンテンツクラスタを T_s とするとき，推薦対象となるコンテンツ c（図 9 における $\overline{B_s} \cap T_o$ に属するコンテンツ）の推薦度 $R(T_s, c)$ を，c が帰属している任意の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 $\mathrm{sim}(T_s, T_{o_i})$ の和によって定義する．

$$R(T_s, c) = \sum_{i=1}^{k} \mathrm{sim}(T_s, T_{o_i}) \tag{5}$$

ただし，類似度が $\mathrm{sim}(T_s, T_{o_i}) < 0$ となる場合においては，式 (4) の仮説検定において p_0 と判定されたと見なし，和に加えないこととする．

既提案法では，1 つのコンテンツクラスタに対して推薦するにあたり，すべての他ユーザのすべてのコンテンツクラスタ間と類似度を算出し，コンテンツの推薦を行った．一方本手法では，すべてのコンテンツクラスタとの類似度算出を行わず，一部のコンテンツクラスタを推定近傍プロファイルとして選択し，それらとの類似度算出のみで可能な限り推薦精度を高めることを目的としている．

© 2009 Information Processing Society of Japan

4.3　推薦精度の定義

今回利用する推薦アルゴリズムは，あるクエリとなるコンテンツクラスタ T_s においてブックマークされていないコンテンツの推薦を行うものである．しかしながら，ある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，ここではすでに帰属関係が示されているコンテンツ集合，すなわち，当該ユーザのブックマーク B_s に含まれるコンテンツ集合に対して推薦度を算出し，推薦結果と，元のコンテンツクラスタに含まれていたコンテンツとの一致度を調べること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)　検証するクエリコンテンツクラスタ T_s を選択し，T_s に帰属しているコンテンツを正解集合 X，T_s に帰属していないコンテンツ（$B_s \cap \overline{T_s}$）を不正解集合 x とする．

(2)　B_s から検証するコンテンツ C_c を 1 つ選択し，T_s および B_s から C_c を除いたデータ T'_s，B'_s と，推定近傍プロファイルであるコンテンツクラスタ T_{o_i}（$i = 1, 2, 3, \cdots$）との間で類似度 $\mathrm{sim}(T'_s, T_{o_i})$ を算出する．

(3)　算出した類似度をもとに C_c の推薦度を算出する．

(4)　(2)–(3) を繰り返し，閾値 Th を超えるものを推薦集合 R，それ未満を非推薦集合 r とする．

(5)　X，x，R，r より recall および precision を算出する．

recall，precision は次式で与える．

$$\mathrm{recall} = \frac{|R \cap X|}{|X|} \tag{6}$$

$$\mathrm{precision} = \frac{|R \cap X|}{|R|} \tag{7}$$

また，推薦精度の検証基準として F-measure を用いる．F-measure は次の式で与えられる．

$$\text{F-measure} = \frac{2 \cdot \mathrm{recall} \cdot \mathrm{precision}}{\mathrm{recall} + \mathrm{precision}} \tag{8}$$

以下の実験では全 300 個のコンテンツクラスタについて検証し，その平均を算出した．また，閾値は $Th = 10$ とした．

4.4　プロファイル選択ネットワークの作成法

本手法で想定するプロファイル選択ネットワークは，各ノードで高い類似度のノードとリンクしている状態が適している．そこで，本研究では，プロファイル選択ネットワークが最適に成長することを想定している．プロファイル間の類似度がまったく計算されていない状態を初期状態と考え，まず，各ノードからランダムに k 個のノードにリンクを張り，初期のプロファイル選択ネットワークを作成する．その後，推薦計算における類似度算出の結果などを利用して，各ノードが自律的に類似度の高いノードへとリンクを張り替える．この操作が繰り返されることにより，プロファイル選択ネットワークが最適化され，推定近傍プロファイルによる推薦精度が高まっていく．

本稿ではその基礎検討として，あらゆるネットワークにおける比較数および推薦精度の評価を行うことで本提案の有効性を示す．ここでは今回実験に用いるプロファイル選択ネットワークの作成方法について説明する．

4.4.1　ランダムネットワーク

各ノードが，自ノードをのぞくすべてのノードの中から k 個のノードをランダムに選択してリンクしたネットワークを，ここでは k–ランダムネットワークと定義する．k が大きければ大きいほど，事前の情報を多く知っているため，少ない比較数でも推薦精度を高めることができると考えられる．また，比較数が全ノード数 −1 のときは，データベース中のプロファイルすべてを利用した従来の協調フィルタリング手法と一致する．

4.4.2　偏りネットワーク

各ノードが，自ノードと類似度の高い上位 k 個のノードとリンクしたネットワークを，ここでは k–偏りネットワークと定義する．これは，自律成長後のネットワークを表現している．

4.5　実 験 結 果

4.5.1　比 較 手 法

本実験においては，提案探索方式を proposed choice，ランダム探索方式を random from NW としている．また，近傍プロファイルを用いた推薦の上界として，全プロファイルから上位 k 件のプロファイルを用いた推薦を**類似度上位による推薦 optimal** とする．

さらに，プロファイル選択ネットワークとは独立して，全プロファイルから単純に無作為抽出した推薦を，**無作為抽出による推薦** random from all とする．

4.5.2　ランダムネットワーク

k–ランダムネットワークにおいて，クエリ数を変化させたときの F-measure を求める実験を 50 回繰り返し，その平均値を算出した．**図 10**，**図 11**，**図 12** に，k = 20，100，299 としたときの結果をそれぞれ示す．今回は全体ノード数が 300 であるので，すべてのプロファイル間類似度に対して，図 10 のときは全体の約 1/15 が既知，図 11 のときは全体の約 1/3 が既知，図 12 のときはすべてが既知となっている．

© 2009 Information Processing Society of Japan

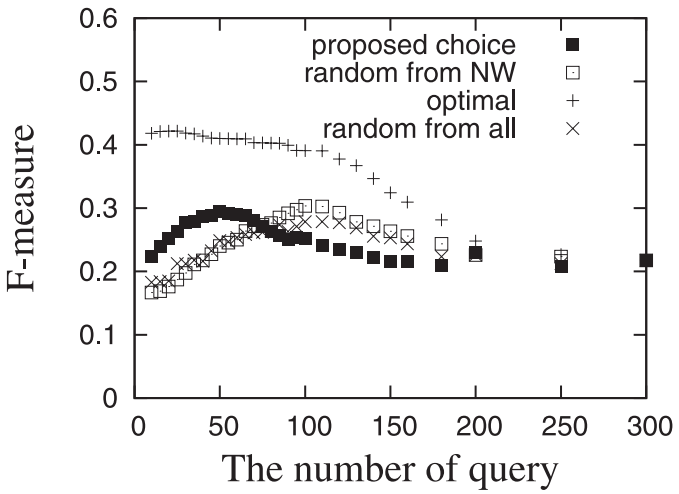

図 10　20–ランダムネットワークの推薦結果
Fig. 10　The result of evaluation in 20-random network.

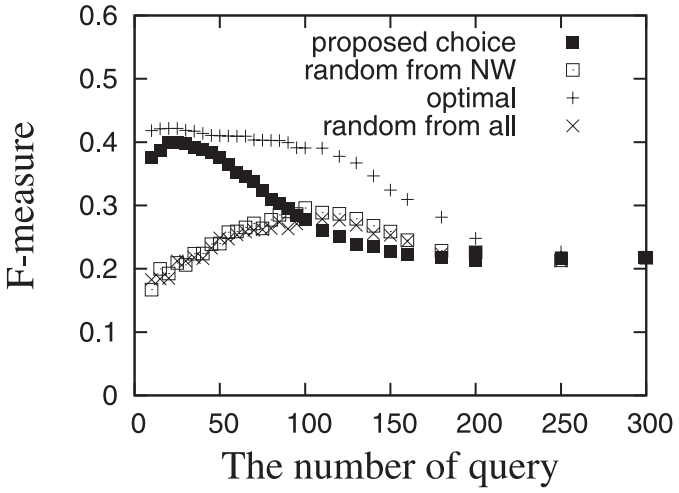

図 11　100–ランダムネットワークの推薦結果
Fig. 11　The result of evaluation in 100-random network.

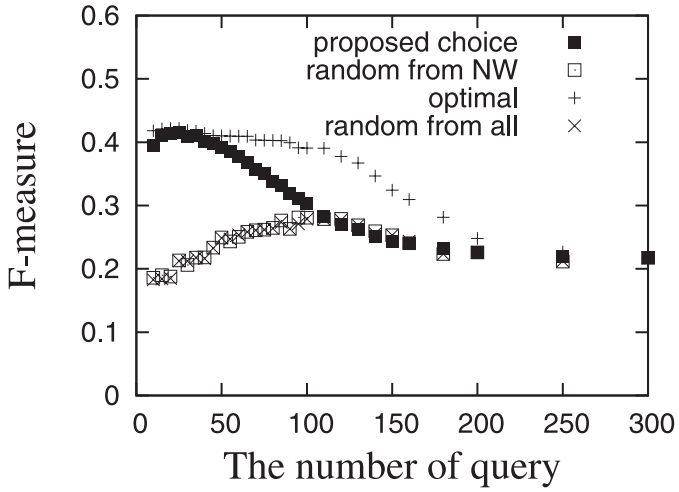

図 12　299–ランダムネットワークの推薦結果
Fig. 12　The result of evaluation in 299-random network.

図 11，図 12 を見ると，特に比較数が小さいときに，提案探索方式がランダム探索方式よりも有用であることが分かる．一方，図 10 の結果において，比較数が小さい領域では提案探索方式の有効性が示されているものの，比較数が大きい領域では提案探索手法の性能がランダム探索方式を若干下回っている．

4.5.3　偏りがあるネットワーク

k–偏りネットワークにおいて，クエリ数を変化させたときの F-measure を求める実験を 50 回繰り返し，その平均値を算出した．**図 13** に $k = 20$ のときの結果を，**図 14** に $k = 100$ のときの結果を示す．これらの結果から，比較数が小さいときは提案探索方式が有効であるが，中程度（70〜150）においてはランダム探索方式のほうが有効となりうるといえる．

加えて，キャッシュ量と推薦精度の関係を調べるため，提案探索方式利用時における 20–偏りネットワーク（20–Biased NW），100–偏りネットワーク（100–Biased NW），299–偏りネットワーク（299–Biased NW，299–ランダムネットワークと一致）の結果を**図 15** に示す．これらの図より，20–偏りネットワークにおける提案探索方式の性能の劣化は顕著なものではないことが分かる．

© 2009 Information Processing Society of Japan

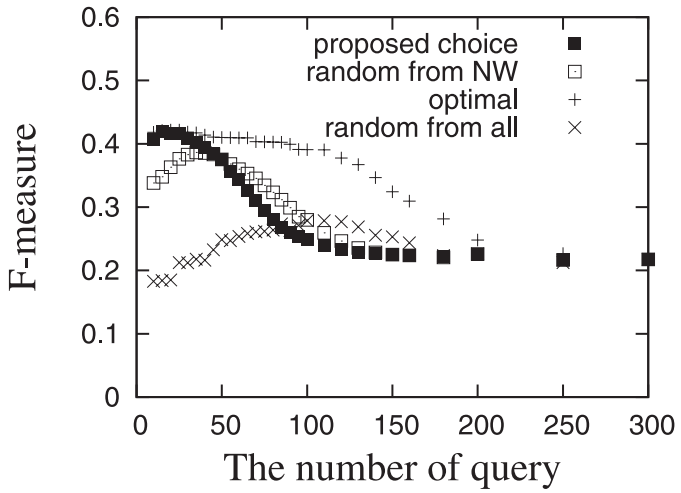

図 13　20–偏りネットワークの推薦結果
Fig. 13　The result of evaluation in 20-biased network.

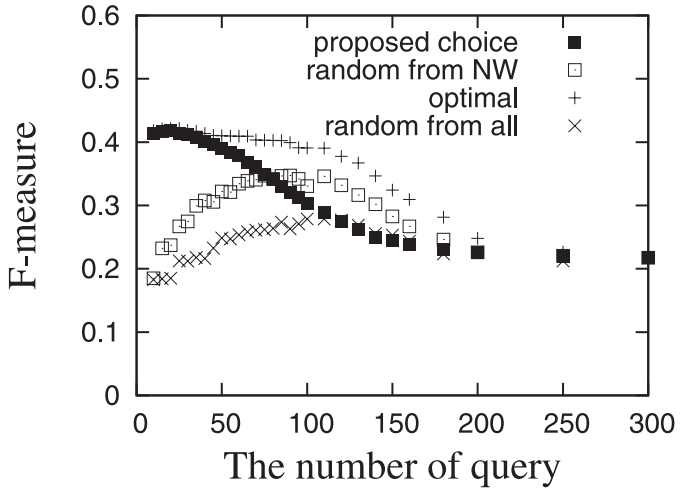

図 14　100–偏りネットワークの推薦結果
Fig. 14　The result of evaluation in 100-biased network.

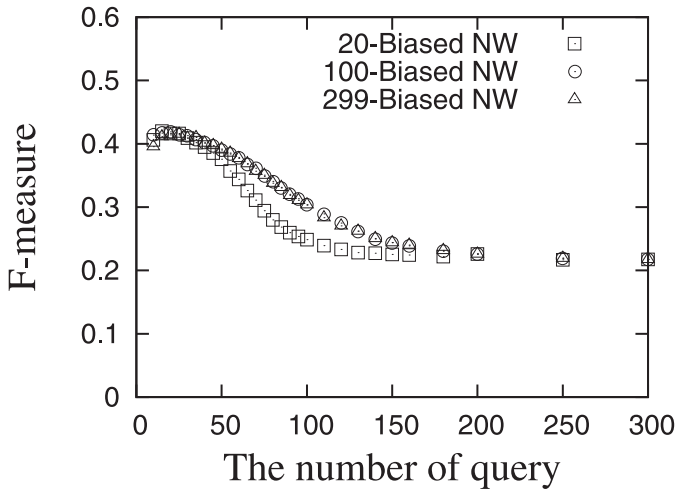

図 15　キャッシュ量と推薦精度の関係
Fig. 15　Relationship between cache and recommendation accuracy.

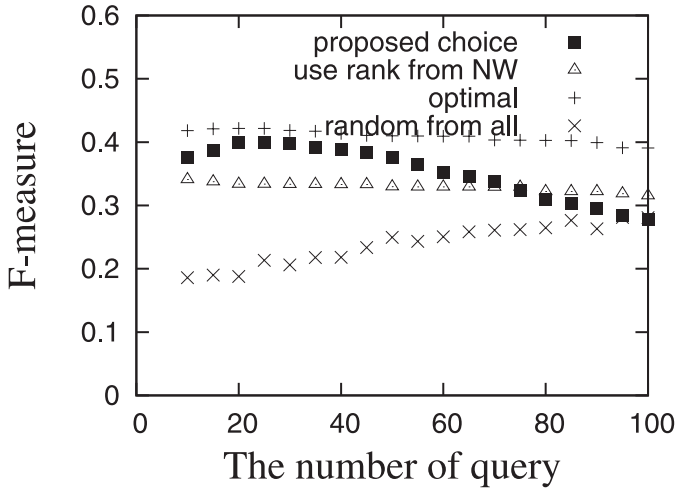

図 16　推移関係利用と非利用の比較
Fig. 16　Comparison between using transitive law and not using transitive law.

© 2009 Information Processing Society of Japan

4.5.4　推移関係の検証

提案探索方式は類似度の推移関係に着目したものであるが，その実際の効果について検証した．**図 16** は，100–ランダムネットワークに対しての推薦結果であり，‘use rank from NW’ は，100 件のリンク先ノードの中の上位のノードから用いて推薦した，1 ホップのみの情報を用いて推薦したものである．この検証結果より，特に比較数が小さいときに提案探索方式が優れていることが分かり，推移関係に着目して探索することが有効であることが示されている．

5. 考　　察

キャッシュとしてプロファイル間類似度が十分な数だけ得られているとき，提案探索方式は有効に機能することが確認された．図 10，図 11 のように，ネットワークの偏りが不十分なときでも，提案探索方式は限られたキャッシュから近傍プロファイルを推定できており，ランダム探索方式よりも良い結果となっている．

さらに，ネットワークに十分な偏りがある状態を想定した，図 13，図 14 では，**比較数が 50 以下においては**厳密な近傍プロファイルを用いた上界と考えられる，類似度上位による推薦に匹敵する推薦精度を示しており，提案探索手法の有効性が確認できている．

しかしながら一方で，比較数が **50 よりも**大きくなると，**提案探索方式が類似度上位の推薦精度と比較して推薦精度が低下し，**ランダム探索方式と比べて有意差を示さない場合が見られた．これは，近傍になりうるプロファイル数が比較数を下回ってしまい，残りのプロファイルからむやみに選択すると，プロファイル選択の傾向が極端に偏ってしまい，アイテムのカバー率が低くなるためであると考えられる．

そこで，本手法の適用範囲となる比較数 Q について考察する．**提案探索方式は近傍プロファイルを近似的に推定するため，有効な比較数 Q の最大値は（厳密な意味での）近傍プロファイルに含まれているプロファイル数（以下，近傍プロファイルサイズ）と等しくなる．**よって，そのデータベースにおける，各プロファイルごとの厳密な近傍プロファイルサイズを知ることができれば，本手法の厳密な適用範囲を推定することが可能であると考えられる．**たとえば本実験においては，任意のキャッシュにおいて最適値に近い結果が得られる $Q = 50$ 付近が近傍プロファイルサイズであると推定できる．**しかしながら，この近傍プロファイルサイズの推定については，データベースの規模（全プロファイル数や登録されたアイテム数）だけでなく，当該データベース中における全近傍プロファイルサイズの偏りまで考慮する必要があるため容易ではない．また，仮に近傍プロファイルサイズを正確に予測できたとしても，それは Q の最大値を与えるのみであり，何をもって最適値とするのか（たとえば，recall と precision のどちらを重視するのかなど）によっても大きく変わるものと考えられる．以上の理由より，Q の最適値については，現状では各サービスごとの求める用件に基づいて，良い値を発見的に得ることが最も有効な方法であると考えられる．

加えて，図 16 の結果から分かるとおり，偏りがない状態で上位のプロファイルを用いた推薦を行うと推薦精度は低くなってしまうが，提案探索手法はこのような場合でも有効な推薦が可能となっている．以上の結果は，偏りの有無にかかわらず，提案探索方式を用いると良い推薦結果が得られるという，アルゴリズムの堅牢性を示す結果であるといえる．

次に，図 15 のキャッシュを制限したときのネットワークに偏りがあるときの検証においても，キャッシュ制限の緩いときに匹敵する推薦精度を持つことを示した．この検証により，本手法が実際の運用に適ったものであることが示された．

また，プロファイル選択自体における計算量について考えると，1 度の探索においては，比較数回の乱数発生と比較数回分の分岐判定のみであり，1 度につき数十万件のアイテムの共起を確かめるプロファイル比較に比べて十分に小さい．

最後に，プロファイルを限定することによる推薦精度の向上効果について考察する．図 10〜14 に示している optimal の F-measure は，全プロファイルである比較数 299 のときに最小値となる．つまり，近傍プロファイル（上位 k プロファイル（$k < 299$））のみを用いて推薦を行った方が，全プロファイル（299）を用いて推薦を行ったときよりも推薦精度が高いことが分かる．プロファイルの限定というアプローチは，**局所的**な嗜好パターンへの対応を可能とし，単に計算量の削減にとどまらず，協調フィルタリングの性能そのものを向上させうることを示唆している．

これに対して，既存の pLSA 手法[6] などは，**大域的**に均一なモデルを，事前に与えられた確率変数の個数分作成するものであり，嗜好パターンの数を事前に決定する必要がある，局所的な嗜好パターンに対応できない，といった問題がある．神嶌らは文献 14) において，局所性の高いデータに大域的なモデルを適用しても，その特殊性を十分に反映したモデルの構築はできず，いわゆる少数派に属するユーザに対して良い推薦ができないことを示しているが，本稿においてもその傾向が現れたといえる．

ウェブ上のコンテンツおよびユーザの関係は，局所的な Web コミュニティ[11)–13)] に分化していることが知られているが，このウェブ上のコンテンツを対象とする推薦においても同様の傾向が現れているといえる．pLSA 手法は大局的でトップダウンなアプローチをとっているため，その局所に対する推薦精度の低下は免れない．一方本手法では，プロファイル

© 2009 Information Processing Society of Japan

全体をモデル化する手法ではなく，各コミュニティごとに適したモデルを逐次作成できるものであるため，事前にモデルの数を決定する必要がない，少数派に属するユーザ対しても適切な推薦が可能となる，といった特徴があげられる．このように本手法は，計算量の削減だけではなく，局所に対する柔軟な推薦を可能とするため，推薦精度の向上も期待できる．

6. おわりに

本稿では，協調フィルタリングにおける計算量削減のための効率的なプロファイル選択手法について検討した．プロファイル間の類似度の推移関係に着目したプロファイル選択方式として，プロファイル選択ネットワークを想定し，類似度に基づいたホップバイホップのノード探索を行うことによって，推薦精度を高く保ったまま，推薦に利用するプロファイル数の削減に成功した．

今後は探索方式の改善および，本手法と協調フィルタリングを合算したときの推薦システムの計算時間を検証する．今回の探索方式においては，被推薦対象となるノードからのリンクの深さを考慮しなかったが，これを考慮することによる性能改善が得られるか，などを今後検討する必要がある．また，今回はネットワークは事前に与えられている状態を想定したが，キャッシュがない初期状態から，最適なネットワークを構築するためのネットワークの成長手法についても検討し，さらに，今回は疑似的にネットワークを扱ったが，実際の分散環境への適用も行う予定である．

参考文献

1) Goldberg, D., Nichols, D., Oki, B.M. and Terry, D.: Using collaborative filtering to weave an information tapestry, *Comm. ACM*, Vol.35, No.12, pp.61–70 (1992).
2) Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews, *Proc. 1994 Computer Supported Cooperative Work Conference*, pp.175–186 (1994).
3) Sarwar, B.M., Karypis, G., Konstan, J.A. and Riedl, J.: Item-based collaborative filtering recommendation algorithms, *Proc. 10th International World Wide Web Conference* (*WWW10*), pp.285–295 (2001).
4) Herlocker, J.L., Konstan, J.A., Borchers, A. and Riedl, J.: Evaluating collaborative filtering recommender systems, *ACM Trans. Inf. Syst.*, Vol.22, No.1, pp.5–53 (2004).
5) 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，情報処理学会論文誌：データベース，Vol.48, No.SIG20, pp.14–27 (2007).
6) Hofmann, T. and Puzicha, J.: Latent class models for collaborative filtering, *Proc. 16th Int'l Joint Conf. on Artificial Intelligence*, pp.688–693 (1999).
7) Zhang, B., Hsu, M. and Forman, G.: Accurate recasting of parameter estimation algorithms using sufficient statistics for efficient parallel speed-up, *Proc. 4th European Conf. on Principles of Data Mining and Knowledge Discovery*, pp.243–254 (2000).
8) Das, A., Datar, M., Garg, A. and Rajaram, S.: Google News personalization: Scalable online collaborative filtering, *Proc. 16th Int'l Conf. on World Wide Web* (*WWW2007*), pp.271–280 (2007).
9) delicious. http://delicious.com/
10) はてなブックマーク．http://b.hatena.ne.jp/
11) Kleinberg, J.: Hubs, authorities, and communities, *ACM Computing Surveys*, Vol.31, Issue 4es, Article No.5 (1999).
12) Gibson, D., Kleinberg, J. and Raghavan, P.: Inferring Web communities from link topology, *Proc. 9th ACM Conference on Hypertext and Hypermedia* (*HyperText 98*), Pittsburgh, PA, pp.225–234 (June 1998).
13) 野村早恵子，小山　聡，早水哲雄，石田　亨：WEB コミュニティ発見のための HITS アルゴリズムの分析と改善，電子情報通信学会論文誌 D-I，Vol.85-D-I, No.8, pp.741–750 (2002).
14) 神嶌敏弘，赤穂昭太郎：参加システムの嗜好パターンが異なる場合の集団協調フィルタリング，人工知能学会研究会，SIG-FPAI-A702-03 (2007).

(平成 20 年 12 月 19 日受付)
(平成 21 年 4 月 4 日採録)

(担当編集委員　河合 由起子)

佐々木　祥（学生会員）

昭和 55 年生．平成 17 年神奈川大学工学部電気電子情報工学科卒業．平成 19 年東京工業大学大学院理工学研究科集積システム専攻修士課程修了．同年より同大学院集積システム専攻博士課程に在学．情報推薦システムに関心を持つ．

© 2009 Information Processing Society of Japan

宮田　高道（正会員）

昭和 53 年生．平成 13 年富山大学工学部卒業．平成 15 年同大学大学院理工学研究科博士前期課程修了．平成 18 年東京工業大学大学院理工学研究科博士後期課程修了．同年より同大学院集積システム専攻助手．平成 19 年より同助教．画像符号化，画像処理，情報推薦等の研究に従事．博士（工学）．電子情報通信学会，映像情報メディア学会各会員．

稲積　泰宏

昭和 51 年生．平成 10 年富山大学工学部卒業．平成 12 年同大学大学院理工学研究科博士前期課程修了．平成 15 年東京工業大学大学院理工学研究科博士後期課程修了．同年神奈川大学工学部助手．平成 19 年より富山大学大学院理工学研究部（工学）講師．画像符号化，画像処理の研究に従事．博士（工学）．IEEE，電子情報通信学会，映像情報メディア学会，画像電子学会各会員．

小林　亜樹（正会員）

昭和 46 年生．平成 7 年東京工業大学工学部情報工学科卒業．平成 9 年同大学大学院理工学研究科電気・電子工学専攻修士課程修了．平成 12 年同大学院博士課程修了．同年同大学院集積システム専攻助手．平成 18 年独立行政法人メディア教育開発センター助教授．平成 19 年同准教授．平成 20 年より工学院大学工学部情報通信工学科准教授．画像検索，ネットワーク情報検索，コンテンツ配信の研究に従事．博士（工学）．電子情報通信学会，映像情報メディア学会，日本データベース学会各会員．

酒井　善則（正会員）

昭和 21 年生．昭和 49 年東京大学大学院工学系研究科電気工学専攻博士課程修了．同年電電公社入社，電気通信研究所勤務．ディジタル伝送，マルチメディア通信会議の研究開発に従事．昭和 62 年東京工業大学助教授．平成 2 年より同教授．映像情報伝送，画像情報検索，情報ネットワークの研究に従事．工学博士．平成 13 年電子情報通信学会業績賞受賞．IEEE-COM，CS 会員．

© 2009 Information Processing Society of Japan