ブックマークの階層構造を考慮した協調フィルタリングによるWebページの推薦手法

A Web Page Recommendation Method with Collaborative Filtering Using User's Bookmark Hierarchy

佐保田 圭介♡ 波多野 賢治♢
宮崎 純♢ 植村 俊亮♢

Keisuke SAHODA Kenji HATANO
Jun MIYAZAKI Shunsuke UEMURA

本論文では，協調フィルタリングにおける共有情報として，利用者のブックマークに登録されている Web ページ内の単語の出現頻度だけではなく，ブックマークの階層構造も利用し，それに基づいて Web ページの推薦を行う手法を提案する．提案手法によって，単語の出現頻度だけに基づいた Web ページ推薦手法に比べ，利用者の嗜好をより反映した Web ページの推薦が可能となった．

In this paper, we propose an Web page recommendation system with collaborative filtering technique based on both term frequency in bookmarked Web pages and hierarchical structure of user's bookmark. Proposed method enables to recommend Web pages meeting user's preference compared with other methods based on only term frequency in bookmarked Web pages.

1. はじめに

インターネット上では，Web ページを用いて誰もが情報を公開し発信することができる．そのため，Web ページの数は日々増加の一途をたどり，2004 年 2 月の時点で，Google[1] に登録されている Web ページ数は 42 億ページ以上となっている．そこで，膨大な量の Web ページから，利用者が必要なページを効率よく見つけるための手段として，様々な Web 検索エンジンが公開され，広く一般に利用されている．しかし，この Web 検索エンジンには，検索結果が利用者の嗜好に依存しない，すなわち，同じ問合せを検索エンジンに入力した場合，誰が問合せを行っても同じ検索結果が返されるという問題点がある．

そこで近年，利用者の嗜好を検索結果に反映させる方法として協調フィルタリングが Web ページの検索に応用され始めている．協調フィルタリングとは，利用者間で嗜好を共有し，嗜好の似た利用者が有用な Web ページと評価したページを推薦する手法である．協調フィルタリングでは，類似した嗜好の利用者の特定のために，共有情報が重要になるが，Web ページ検索という応用分野では，利用者の嗜好が顕著に反映された情報であるという理由でブックマークを共有情報として利用することが多い [5, 8].

本論文では，利用者独自に作成しているブックマークを利用し，ブックマークに登録されている Web ページ内の単語の出現頻度を協調フィルタリングにおける共有情報として利用するだけではなく，ブックマークの階層構造も利用し，それに基づいて Web ページの推薦を行う手法を提案する．つまり，類似した嗜好を持つ利用者を特定するために，ブックマークに登録されている個々の Web ページの類似度の他に，ブックマークの階層構造から得られるフォルダの類似度，構造の類似度を利用する．ブックマークの階層構造から得られる類似度は，ブックマーク全体での利用者の嗜好を評価するものではなく，ブックマークの一部分に注視し，利用者の部分的な嗜好の類似性を判定するという特徴を持つ．

以下，本論文では，ブックマーク中の Web ページをブックマークページ，利用者がブックマークに加えた Web ページを追加ブックマークページ，また，提案システムを利用し，推薦を受ける利用者を推薦情報享受者，システムにブックマーク情報を提供し，共有情報を提供する利用者をブックマーク提供者と定義し，議論を進める．

2. 関連研究

森らは，現在推薦情報享受者が閲覧している Web ページに対して，ブックマーク提供者のブックマークページを推薦するシステム，ブックマークエージェントを開発した [5]．このシステムは，現在閲覧中の Web ページから推薦情報を生成するため，推薦情報享受者の過去の嗜好に依らない推薦ができるという特徴を持つ．

また，Rucker らは，推薦情報享受者のあるブックマークフォルダ内におけるブックマークページが，ブックマーク提供者のブックマークフォルダ内におけるブックマークページといくつ一致しているかによってフォルダの類似度を定義し，類似度の最も高いフォルダ内に存在するブックマークページを推薦するシステム Siteseer を開発した [8].

以上のシステムでは，協調フィルタリングの共有情報としてブックマークページの内容を利用しているだけでブックマークの階層構造を利用していない．また，ブックマークを深さ 1 の幅が広い木として扱ったり，フォルダ構造を一階層しか利用していないため，ブックマークを十分に活用しているとは言えず Web ページの推薦精度が問題となる．特に後者の研究は，ブックマークのフォルダに含まれる Web ページが多いほど推薦フォルダになり得るため，ブックマークページが少ないフォルダに有益なブックマークページが存在する場合は，そうしたページは推薦されにくく，推薦精度が悪くなるという問題が生じる．

3. 提案手法

3.1 システム概要

本論文で提案するシステムは，推薦情報享受者が Web ブラウズ中に興味を持った Web ページをブックマークに加えたとき，この追加ブックマークページが推薦情報享受者のブックマークのどの部分に追加されるか，また，推薦情報享受者のブックマークの内容および階層構造を用いて，追加ブックマークページに類似するブックマーク提供者のブックマークページをランキング形式で推薦するシステムである (図 1 参照).

Web ページの推薦は，以下の手順で行われる．

1. 推薦情報享受者の追加ブックマークページ α とブックマーク提供者のブックマークページ β 間の Web ページの類似度 R をコサイン相関値を用いて算出する．
2. 追加ブックマークページ α を含む全祖先フォルダとブックマークページ β を含む全祖先フォルダ間のフォルダの類似度

♡ 学生会員 奈良先端科学技術大学院大学情報科学研究科，現在，日本電気株式会社 keisuk-s@is.naist.jp
♢ 正会員 奈良先端科学技術大学院大学情報科学研究科
{hatano,miyazaki,uemura}@is.naist.jp
[1] http://www.google.com/

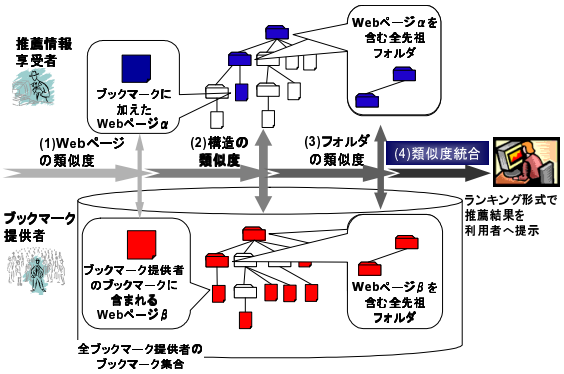

図 1: システム概要

Fig. 1: System Overview

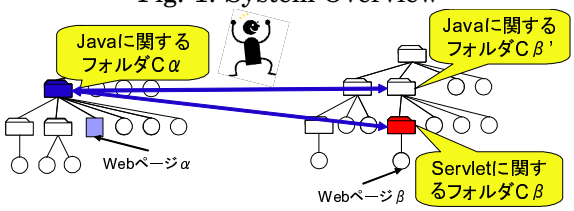

図 2: フォルダの特徴の比較

Fig. 2: Comparison of Folders' Features

のうち，最も類似度が高いものをフォルダ間の類似度 R' として算出する (3.2.1 節で詳述).

3. 推薦情報享受者のブックマークとブックマーク提供者のブックマークの構造の類似度 R'' を edit-distance を元に算出する (3.2.2 節で詳述).
4. 上記 1 ～ 3 で計算された類似度を正規化し評価統合を行う．そして，統合結果を元に推薦情報享受者に推薦結果を提示する (3.3 節で詳述).

3.2 二つのブックマーク構造特徴量

ブックマークの類似判定には，ブックマーク全体の階層構造の他に，ブックマークの一部分に注視しその一部分のブックマークフォルダ内に含まれているブックマークページの内容を解析する必要がある．本節では，利用者の嗜好を分析するために必要なブックマークの構造特徴量を考え，それに基づいた評価方法について説明する．

3.2.1 フォルダの類似度による評価

ブックマークフォルダの中に含まれているブックマークページを，ブックマークの階層構造を利用して解析することで，利用者の嗜好を分析することが可能である．つまり，推薦情報享受者の追加ブックマークページ α とブックマーク提供者のブックマークページ β がどのようなフォルダをたどって分類されているかを比較することで，利用者間の嗜好の類似性判定が可能となる．しかし，推薦情報享受者とブックマーク提供者のブックマーク間には構造上の大きな違いがあり，単純にフォルダ内のブックマークページを比較するだけでは，フォルダ間の類似度を算出することができない．

我々が採用した算出法を，図 2 を用いて具体的に説明する．追加ブックマークページ α を含むフォルダを C_α，ブックマークページ β を含むフォルダを C_β，ブックマークページ β のひとつ上の階層のフォルダを $C_{\beta'}$ とする．ブックマークは深層にいくにしたがって内容が特定されていくことを考えると，図 2 の例では，ブックマーク構造の違いにより，フォルダ C_α とフォルダ C_β の類似度は，フォルダ C_α とフォルダ $C_{\beta'}$ の類似度より低いと考えられる．したがって，我々はフォルダの類似度を正確に計算するために，ブックマークのルートフォルダから追加ブックマークページ α，ブックマークページ β が含まれているフォルダまでのすべての中間フォルダ対して類似度の計算を行い，その最大値をフォルダの類似度とすることで，ブックマークの構造上の違いを吸収できると考えた．フォルダの類似度の算出方法は以下のとおりである．

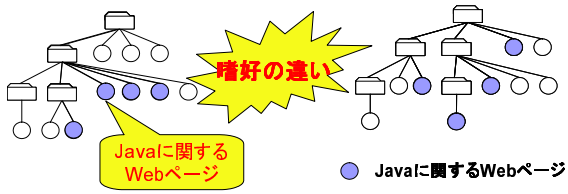

図 3: 利用者間での嗜好の違い

Fig. 3: Difference of User's Preference

フォルダの類似度の算出 フォルダの特徴ベクトルの計算には，tf-idf 法の応用として提案された tf-icf 法 [1] を利用する．tf-icf 法で単語の重みを計算することで，ブックマークフォルダを反映した単語の重み付けを行い，ブックマークフォルダの特徴ベクトルを生成できる．ブックマーク提供者のブックマーク集合に含まれる総フォルダ数 n，フォルダ c に含まれる単語 t^c と単語 t^c を含むブックマーク内のフォルダ数を f_{t^c}，フォルダ c に含まれる単語 t^c の出現頻度を $f^c_{t^c}$ とすると，フォルダ c における単語 t^c の重み $w^c_{t^c}$ は以下のように定義される．

$$w^c_{t^c} = (1 + \log_e f^c_{t^c}) \log_e (1 + \frac{n}{f_{t^c}}) \quad (1)$$

また，フォルダに含まれる単語の数 M とした時，フォルダの特徴ベクトル $\boldsymbol{V}^c$ は以下のように定義される．

$$\boldsymbol{V}^c = (w^c_{t^c_1}, w^c_{t^c_2}, \cdots, w^c_{t^c_M}) \quad (2)$$

最終的に，比較対象フォルダ g の特徴ベクトルを $\boldsymbol{V}^g$ とすれば，フォルダ c と g 間の類似度 R' を次式のようにコサイン相関値から計算できる．

$$R' = sim(\boldsymbol{V}^c, \boldsymbol{V}^g) = \frac{\boldsymbol{V}^c \cdot \boldsymbol{V}^g}{|\boldsymbol{V}^c||\boldsymbol{V}^g|} \quad (3)$$

3.2.2 構造の類似度による評価

ブックマークの管理は利用者に任されているため，利用者は独自の構造でブックマークを管理する．図 3 の例では，Java に関する Web ページでも，利用者によって情報組織化の方法に違いがあるため，異なるブックマークの階層構造をとることがわかる．この階層構造の違いを比較することができれば，利用者間の嗜好の類似性判定に利用することが可能となる．本研究では，ブックマークページの組織化法の類似した利用者からの推薦を受けることが重要となるため，ブックマーク構造の類似度 R'' を定義する．つまり，利用者のブックマーク全体の構造ではなく，追加ブックマークページ α に類似した Web ページを同じように組織化しているブックマーク提供者から Web ページ推薦を受けることを目的とする．

図 4 は利用者間で異なった嗜好で作成された階層構造を持つブックマークの類似性判定手順について示したものである．推薦情報享受者によって Web ページが加えられたブックマーク 1) とブックマーク提供者のブックマーク 2) から，追加ブックマークページ α とある閾値以上の類似度を持った Web ページ r のみを残し，閾値以下の類似度の Web ページを削除するとそれぞれ 3), 4) となる．3), 4) とを比較することで，推薦情報享受者の追加ブックマークページ α に類似しているブックマークページがどのように組織化されているのかが判定でき，ブックマーク構造間の類似性判定を利用者の嗜好の類似性判定に利用することができる．

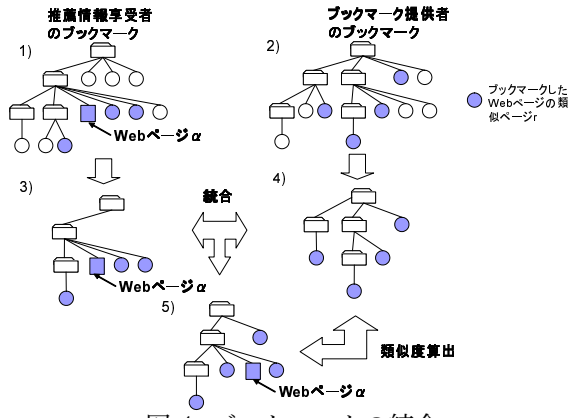

図 4: ブックマークの統合

Fig. 4: Integration of Bookmarks

得られたブックマークの構造間の類似度の判定には edit-distance [6] を用いた．しかし，3) と 4) の構造を比較すると利用者間でブックマークの構造が大幅に異なるため，利用者の嗜好をうまく反映した類似判定を行うことができない．そこで，ブックマーク同士を統合し，中間的なブックマーク構造 5) を作成し，中間的なブックマーク構造と 4) を比較し，最終的に構造の類似度 R'' を算出した．edit-distance とは，任意の文字列 p に対して挿入，削除，置換のいずれかを適用して，任意の文字列 q を求めるための最小コストの算出手法である．ブックマークを深さ優先探索で走査し，ラベル付けを行い，正規化したものを文字列で表すが，その文字列間の類似度に edit-distance で算出された値を用いている．算出された値が大きいほど，構造の類似度が低いことを表している[2]．

一方，ブックマークの統合には，データ統合の研究分野の技術であるメディエータと呼ばれるプログラムを利用した．メディエータとは，問合せの再構築や配布を行ったり，結果データをフィルタリング，統合，処理したりすることにより，新しい，より密度の高い，より関連性のある情報を生成するように，問合せや結果に対して付加価値の処理を行う手法である [7]．メディエータにはさまざまな手法 [2] があるが，本論文では XML のデータを統合できる XML Data Mediator[3] を利用した．

最終的に，ブックマーク構造の類似度 R'' は，edit-distance によって算出されたコストが γ だったとき，以下のように定義される．

$$R'' = \frac{1}{\gamma + 1} \tag{4}$$

3.3 評価統合

提案する推薦手法では，最終的に推薦されるべき Web ページをランキング表示するために，これまでに算出された類似度同士を統合する必要がある．情報検索の分野では二つの値を統合するための関数として様々な関数が用いられており，例えば最大値や最小値，相加平均などが有名である．ところが，これらの関数の選択によって推薦手法の有効性に差が出ることが判明している [9]．Montague らの報告では，文書検索において，SUM を使った正規化とそれぞれの類似度の和を表す CombSUM を使った評価統合を用いた場合，高い検索精度になることが実験で示されている [4]．そこで，我々はこの二つの関数を使用し正規化と評価統合を行う．

正規化前の類似度 $S_i(O_j)$，ブックマーク中の Web ページ数 N の場合，SUM を用いて正規化を行った後の類似度 $S'_i(O_j)$ を以下の式で定義できる．

$$S'_i(O_j) = \frac{S_i(O_j)}{\sum_{j=1}^{N} S_i(O_j)} \tag{5}$$

さらに，計算された各類似度 $S'_i(O_j)$ を CombSUM によって評価統合する場合，評価統合後の類似度 $S''(O_j)$ は以下の式で計算される．

$$S''(O_j) = \sum_{i=1}^{N} S'_i(O_j) \tag{6}$$

表 1: 各利用者のブックマークの詳細

Table 1: Statistics of Collected Bookmarks

推薦情報享受者数	33 人
総フォルダ数	1,724 個
デッドリンクではない Web ページ数	9,201 ページ
フォルダ数の平均	52.24
フォルダ数の標準偏差	62.00
ブックマークの深さの平均	3.39
ブックマークの深さの標準偏差	0.555
ブックマーク中ページの平均	252.24
ブックマークページの標準偏差	352.13

4. 評価実験

本章では，3 章で提案した Web ページ内の単語の出現頻度だけではなく，ブックマークの階層構造も協調フィルタリングにおける共有情報として利用する Web ページ推薦手法の有効性を評価するために評価実験を行った．

4.1 実験環境

情報検索の分野では検索精度を評価するため，TREC テストコレクション[4] が存在する．しかし，本手法の有効性を示す実験では，利用者が独自に作成しているブックマークが必要となるため，既存のテストコレクションを使用できない．そこで，我々の想定に合うようなテストコレクションを独自に構築し，実験を行った．具体的には，Web 上ではブックマークを公開している利用者が多いため，それらのファイルを人手で収集しブックマーク提供者のブックマークとして利用した．公開されたブックマークはフォルダ名，URL の情報を持っており，またブックマークの階層構造も存在している．収集したブックマークの詳細を表 1 に示す．

4.2 評価方法

実験の評価は以下のように行った．

1. 被験者が Web ブラウズ中にある Web ページをブックマークに加えることを想定し，第一著者の友人である 3 人 の被験者によって，一つずつ合計三つの追加ブックマークページについて，適合 Web ページの組を作成する．
2. 追加ブックマークページを提案手法を実装したシステムに入力し，ランキング形式で推薦されたブックマークページを得る．
3. 三つの追加ブックマークページに対して，ランキング形式で推薦されたブックマークページ群と 1. で作成された適合 Web ページを比較し，再現率- 適合率グラフを描き，検索精度の評価を行う．

4.3 実験結果

本実験では，以下の四つの場合を比較することで，協調フィルタリングにおける共有情報として利用者のブックマークに登録されている Web ページ内の単語の出現頻度だけを利用した Web ページ推薦手法との比較を行う．

[2]実際の算出には GSpell (`http://specialist.nlm.nih.gov/GSpell.html`) の Java API を利用した．

[3]XML Data Mediator: `http://www.alphaworks.ibm.com/tech/`

[4]National Institute of Standards and Technology: Text REtrieval Conference, `http://trec.nist.gov/`

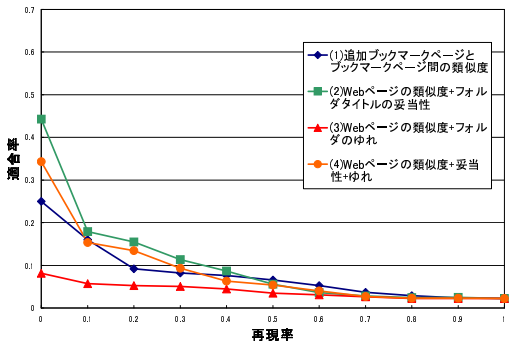

図 5: 実験結果の再現率-適合率グラフ

Fig. 5: Recall-Precision Curves of Each Method

1. ブックマークページと追加ブックマークページ間の類似度のみ
2. 1. の類似度とフォルダの類似度を評価統合したもの
3. 1. の類似度と構造の類似度を評価統合したもの
4. 1. の類似度，フォルダの類似度，構造の類似度を評価統合したもの

各実験の再現率-適合率グラフを図 5 に示す．この結果より，ブックマークページと追加ブックマークページ間の類似度とブックマーク構造の類似度を評価統合した推薦手法が，ブックマークページと追加ブックマークページ間の類似度だけを考慮した場合に比べ高い精度で推薦を行うことできることがわかる．

4.4 考察

我々は，ブックマークページの間の類似度，フォルダの類似度，構造の類似度三つを評価統合した手法が最もよくなると予想していたが，このような結果となってしまった理由として，以下のような問題があったためであると考えられる．

- フォルダの類似度の問題点
  tf-icf 法のよるフォルダの特徴付けには，フォルダ内に含まれるブックマークページが少ないフォルダ同士の類似度の値が高くなるという問題点があることが実験結果を分析することにより判明した．この問題はブックマークに含まれるブックマークページが少ないブックマーク提供者と推薦情報享受者の嗜好が似ている場合に起こる問題であるため，tf-icf 法とは異なるフォルダの特徴づけ手法を検討する必要がある．
- ブックマークページの類似度の問題点
  図 5 を見る限り，ブックマークページの類似度だけで Web ページの推薦を行った場合の推薦精度が低いように思われる．ブックマークページの重み付けは，文書検索や Web 検索の分野で使用されている tf-idf 法を用いたが，今回のように単語数が少ない文書は想定外であることに起因しているように思われる．したがって，含まれている単語が少ないブックマークページに対して，何らかの基準を設けて，ブックマークページとして扱うか否かを判定する機構が必要である．

しかし，単語の出現頻度に基づいたブックマークページの類似度だけを用いた場合に比べ，どの手法も再現率-適合率グラフは上部にプロットすることができ，ブックマークの階層構造をも利用した Web ページの推薦手法の方が，推薦情報享受者と嗜好の類似したブックマーク提供者からの Web ページ推薦を受けることが可能であることがわかった．

5. まとめ

本論文では，協調フィルタリングで共有情報として利用者のブックマークに登録されている Web ページ内の単語の出現頻度だけではなく，ブックマークの階層構造をも利用し，それに基づいて Web ページの推薦を行う手法を提案した．また，評価実験を行った結果，構造の類似度を加味した方がより精度のよい Web ページの推薦が行えることが判明した．

本提案の利点は，推薦情報享受者が独自に作成しているブックマークを利用することで，嗜好に合った Web ページの推薦が受けられるだけではなく，ブックマークの一部分に着目して推薦すべき Web ページを決定しているため，推薦情報享受者の嗜好の変化にも対応できる点である．

今後の課題は，4 章で述べた提案手法が抱えている問題点の解決に加え，他の関連研究，例えば [3] などとの Web ページ推薦精度の比較を行い，本提案の優位性を示すことである．

[謝辞]

本研究の一部は，日本学術振興会および文部科学省科学研究費補助金 (課題番号はそれぞれ 14780325，15017243) の支援によるものである．ここに記して謝意を示す．

[文献]

[1] K. Cho and J. Kim. Automatic Text Categorization on Hierarchical Category Structure by using ICF (Inverted Category Frequency) Weighting. In *Proc. of KISS Conference*, pp. 507–510, 1997.

[2] S. Cluet, C. Delobel, J. Siméon, and K. Smaga. Your Mediators Need Data Conversion! In *Proc. of SIGMOD'98*, pp. 177–188. ACM Press, June 1998.

[3] J. J. Jung, J.-S. Yoon, and G. Jo. Collaborative Information Filtering by using Categorized Bookmarks on the Web. In *Proc. of INAP 2001*, Vol. 2543 of *LNCS*, pp. 237–250. Springer, July 2003.

[4] M. Montague and J. A. Aslam. Relevance Score Normalization for Metasearch. In *Proc. of CIKM 2001*, pp. 427–433. ACM Press, Nov. 2001.

[5] 森幹彦, 山田誠二. ブックマークエージェント: ブックマークの共有による情報検索の支援. 電子情報通信学会論文誌, Vol. J83-D1, No. 5, pp. 487–494, May 2000.

[6] A. Nierman and H. V. Jagadish. Evaluating Structural Similarity in XML Documents. In *Proc. of WebDB 2002*, pp. 61–66, June 2002.

[7] 西尾章治郎, 大田友一, 横田一正, 西田豊明, 佐藤哲司. 情報の共有と統合, 岩波講座 マルチメディア情報学, 第 7 巻. 岩波書店, 1999.

[8] J. Rucker and M. J. Polanco. Siteseer: Personalized Navigation for the Web. *CACM*, Vol. 40, No. 3, pp. 73–75, ACM Press, Mar. 1997.

[9] 鈴木優, 波多野賢治, 吉川正俊, 植村俊亮. 複数のメディアで構成された電子文書の検索手法. 情報処理学会論文誌: データベース, Vol. 42, No. SIG10(TOD11), pp. 11–21, July 2001.

佐保田 圭介　Keisuke Sahoda
奈良先端科学技術大学院大学情報科学研究科博士前期課程修了，現在，日本電気株式会社に勤務．在学中は協調フィルタリングシステムの研究に従事．

波多野 賢治　Kenji Hatano
奈良先端科学技術大学院大学情報科学研究科助手．情報検索システムの研究に従事．情報処理学会，電子情報通信学会，ACM，IEEE CS，各会員．

宮崎 純　Jun Miyazaki
奈良先端科学技術大学院大学情報科学研究科助教授．データベースシステムの研究に従事．情報処理学会，電子情報通信学会，ACM SIGMOD，IEEE CS，各会員．

植村 俊亮　Shunsuke UEMURA
奈良先端科学技術大学院大学情報科学研究科教授．データベースシステムの研究に従事．情報処理学会，電子情報通信学会，各フェロー，IEEE Fellow．