

タグクラスタ多様化による未知性を考慮した推薦手法の提案

藤原 誠[†] 中川 博之[†] 田原 康之[†] 大須賀昭彦[†]

Improving Recommendation Novelty Based on Tag Cluster Diversification

Makoto FUJIWARA[†], Hiroyuki NAKAGAWA[†], Yasuyuki TAHARA[†],
and Akihiko OHSUGA[†]

あらまし 近年推薦システムの研究において、その有用性は予測精度（適合率、再現率）の高さとされていたが、既存の推薦方法では、ユーザは予測精度の高さだけでは満足しないということが報告されている。したがって、今後の推薦システムの課題として、未知性や意外性といった予測精度以外の新たな観点が重要とされている。こうした課題を踏まえ、本論文では推薦リストの未知性という観点に着目し、ユーザの嗜好に合致するが、まだ知らないコンテンツの発見を促す推薦手法を提案する。多面的に付与されたタグの話題単位で推薦リストを多様化することにより、予測精度を維持しつつ、未知性を向上させる。評価実験の結果、推薦リストの約6割に嗜好に合致し、かつ知らないコンテンツが含まれ、未知性の高い推薦が実現できていることを確認した。

キーワード 推薦システム、フォークソノミー、未知性

1. ま え が き

近年、インターネットの急速な普及により、膨大な情報を個人が取得できるようになった。そのため、ユーザが本当に望む情報を取得することが困難になってきており、情報フィルタリングの重要性が高まっている。情報フィルタリングの一つとして推薦システムは、多くの情報からユーザの望む情報をフィルタリングする有用な技術として、様々なサービスに広く用いられ、大きな成功をおさめている [1]。従来の推薦システムの研究において、推薦の有用性は、推薦がユーザの嗜好に合致しているかという予測の正確さとして捉えられ、その評価指標として予測精度（適合率、再現率等）の向上が重視されてきた。しかし、従来の協調フィルタリングやコンテンツベースの推薦方法では、予測精度が高いがゆえに嗜好に合致するが似たようなコンテンツばかり推薦されてしまい、結果的にユーザは予測精度の高さだけは満足しないと報告されている [2]。したがって、今後の推薦システムの課題として、推薦システムの有用性を従来の予測精度という観点だけではなく、ユーザに新たな価値の発見を提供するといった観

点から検討することが重要とされている [3], [4]。例えば、ユーザにとってまだ知らないコンテンツや意識していないが潜在的な嗜好に合致するコンテンツを推薦することにより、推薦システムへの満足度を向上させることができると考えられる。

こうした課題を踏まえ、本論文では推薦の未知性という観点に着目し、ユーザにとって嗜好に合致するが、まだ知らないコンテンツの発見を促す推薦手法の構築を目的とする。本論文における未知性とは、「ユーザの嗜好に合致するが、まだ知られていないコンテンツ」と定義する。一般的に嗜好に合致するコンテンツを推薦することとまだ知らないコンテンツを推薦することとの間には、トレードオフの関係があるとされており [5]、既存研究では未知性を向上させているものの予測精度に関しては低下しがちである。そこで、folksonomy（フォークソノミー）の多面的に付与されたタグに基づき、各話題単位で推薦リストを多様化することで、こうしたトレードオフが解消できるという仮説のもと、未知性を考慮した推薦手法を提案する。folksonomy は、ユーザが情報にタグというキーワードを自由に付与する分類方法で、近年ブックマークを共有する delicious [6] やはてなブックマーク [7]、写真を共有する flickr [8]、動画を共有する Youtube [9] 等のサービスが普及している。従来の taxonomy を用いた分類に比べ、分類が多面的であるという利点から

[†] 電気通信大学大学院情報システム学研究所, 調布市
Graduate School of Information Systems, The University
of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi,
182-8585 Japan

ユーザの新たな発見を支援できる可能性を秘めているが、タグの表記ゆれや多義性などの欠点から生じる予測精度低下のため、folksonomy の利点をこのような観点で活かそうという試みはほとんどなされていない。本論文ではこうした folksonomy 特有の欠点から生じる予測精度低下を階層的クラスタリング及びユーザプロフィールに基づいたタグクラスタの選択・粒度決定により解決し、まだ知らないコンテンツを推薦する。

評価実験では、代表的なソーシャルブックマークサービスである「はてなブックマーク」を題材にする。ソーシャルブックマークは、Web サイト上で複数ユーザのブックマーク（お気に入りの Web ページ）を共有するサービスであり、ユーザはブックマークを登録する際に Web ページに自由に複数種類のタグを付与することができる。また flickr や Youtube などのコンテンツにおいても同様にタグが付与されており、これは一般的なタグとコンテンツの関係といえる。はてなブックマークは 2007 年にサービスが開始され、2011 年 12 月現在でエントリ数約 4000 万、ブックマーク数約 1 億 1000 万という大規模なソーシャルブックマークサービスであり、folksonomy を利用した代表的なサービスといえる。一つのコンテンツ（Web ページ）に付与されるタグが多岐にわたっており、タグの表記ゆれや多義性が見受けられるという点から、本研究が想定する folksonomy を用いたサービスとしての条件に合致している。「はてなブックマーク」の実データを用いて推薦リストの未知性に関する評価実験を実施した結果、推薦リストの約 6 割にユーザがまだ知らないコンテンツが含まれ、提案手法では既存手法より未知性の高い推薦が実現できていることを確認した。

本論文の構成を以下に示す。2. では関連研究として予測精度以外の指標を重視した推薦システム及び folksonomy を用いた推薦システムに関する関連研究について述べる。3. では folksonomy について説明し、4. では、本研究で提案する推薦手法について述べる。5. では、提案手法の有効性を評価するために行った評価実験について述べる。6. では、実験結果に対する考察、7. では結論を述べる。

2. 関連研究

2.1 未知性に関する関連研究

推薦リストにおける未知性とは、一般に「嗜好に合致しているがユーザがまだ知らないコンテンツ」を指す。本論文でも同様に、ユーザの嗜好の範囲内でまだ

知られていないコンテンツを未知性のあるコンテンツと定義する。本論文の未知性に近い概念の研究としては、Wu [10] ら、住元ら [11]、Park ら [12]、Hijikata ら [13]、[14] の研究がある。Wu [10] らは、動画間の類似度を測り、好みのコンテンツと似ているが、人気のないロングテールなものを推薦することで未知性を向上させている。住元ら [11] は、Pixiv [15] を対象に、有力投稿者を定義し、その投稿者がブックマークしていて、かつロングテールなコンテンツを推薦し、未知性を向上させている。また Park ら [12] は、ロングテールのテール部分に位置するコンテンツをクラスタリングし、推薦精度の向上を行っている。これらの研究はいずれもロングテール部分に属するコンテンツから嗜好に合致しそうなものを選ぶことにより未知なコンテンツの推薦を促している。これらの研究ではロングテール部分のコンテンツに着目しているため、コンテンツ自体の質が低下する懸念があり、特にソーシャルブックマークのようなコンテンツの質に差があるサービスを対象とした場合、その懸念が特に顕著に表れると考えられる。また Hijikata ら [13]、[14] は、ユーザから明示的に与えられる、コンテンツに対する既知、不既知のプロファイルを用いて、ユーザから推薦されるコンテンツが既知である確率を計算し、推薦リストの未知性を向上させている。しかし、知らないコンテンツに対して「知らない」という情報を明示的に獲得することは難しいと考えられるため、ユーザの入力コストという観点では課題が残る。

2.2 推薦リストの多様化に関する関連研究

推薦リストの多様化によって未知なコンテンツや意外なコンテンツの推薦を促す研究として Zeigler ら [2]、[16]、小川ら [17] の研究がある。Zeigler ら [2]、[16] は、taxonomy によりコンテンツの類似性を求め、推薦リスト内の全コンテンツの類似性を “Intralist Similarity” という指標で定義している。類似度の高いコンテンツを推薦リスト内に表れにくくすることで、多様性を高める手法を提案し、適合率や再現率は低下するものの、多様性が向上すると同時にユーザの満足度も向上することを示している。本論文では、ユーザの満足度における一要素として推薦リストの未知性に着目している。小川ら [17] は、ユーザの共評価値という従来の taxonomy と異なる観点でアイテムをクラスタリングし、このクラスタリングしたトピックに基づいて多様化することにより、推薦リストの未知性を向上させている。この研究では未知性 (Novelty) を「今まで知

らなかったが、推薦されて興味をもったアイテム」と定義しており、これは一般的に意外性 (Serendipity) という概念に近い。またこれらの研究は本や映画のDVDを対象に、あらかじめコンテンツをある基準に従い一意に分類しておき、そのカテゴリーやトピックを横断することで新たな発見を促している。しかしながら、ソーシャルブックマークのような多様な観点で分類されるようなコンテンツを対象にした場合、個々のコンテンツが複数の分類語 (タグ) に属していることや個々ユーザによって分類の観点が異なることから、必ずしも一意にカテゴリー分けできるわけではない。したがって、先行研究のように一意な分類に基づいて多様化した推薦を行った場合、ユーザが意図している嗜好の範囲を超えてしまい、未知なコンテンツは多く含むが嗜好に合致しないコンテンツが推薦されてしまう可能性が高いと考えられる。

2.3 folksonomy を用いた推薦システムの関連研究

folksonomy のタグ情報を利用した推薦手法の研究には、Guy ら [18], Niwa ら [19], [20], Shepitsen ら [21] の研究がある。Guy [18] らはユーザベースとタグベースを組み合わせた協調フィルタリングにより様々な種類のコンテンツをパーソナル化して推薦することで予測精度の向上を図っている。Niwa [19], [20] らは、クラスタリングによってタグの表記ゆれという問題を解消している。Shepitsen [21] らは、ユーザに明示的にタグを選んでもらい、それに基づいてタグクラスタの粒度を決定することによりパーソナル化した推薦システムを提案している。これらの研究は、folksonomy の欠点であるタグの表記ゆれ、多義性、及び付与頻度の低いタグの存在といった問題に対し、クラスタリングすることや既存手法と組み合わせることにより解決を試みており、推薦システムの予測精度を向上させている。しかしながら、これらの研究は推薦システムの予測精度のみを高めることを目的とした研究であり、未知性など予測精度以外の観点については重視されていない。

3. folksonomy

3.1 folksonomy とは

folksonomy とは、ユーザが自らインターネット上の情報に複数のタグというキーワードを自由に付与する分類方法である。人々を意味する “folks” と分類法である “taxonomy” に由来している。また folksonomy

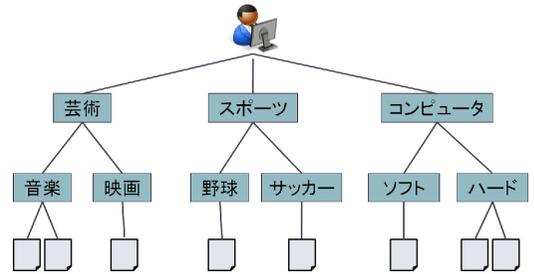


図 1 taxonomy による分類
Fig. 1 A classification based on taxonomy.

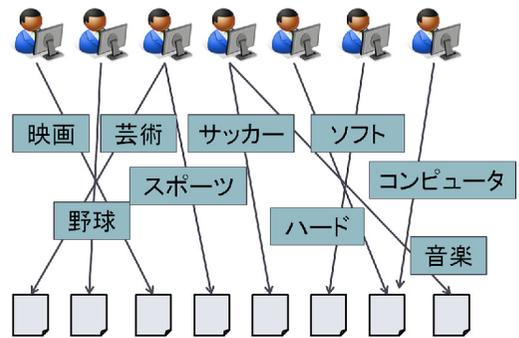


図 2 folksonomy による分類
Fig. 2 A classification based on folksonomy.

は伝統的な分類方法である taxonomy と対比されることが多い。従来の taxonomy はあらかじめ決定された分類木に基づいて、一意に分類するトップダウンな分類方法といえる (図 1)。それに対し、folksonomy は個々のエンドユーザがそれぞれタグ付けを行うことにより分類するボトムアップな分類方法といえる [22] (図 2)。

3.2 folksonomy の特徴

folksonomy を従来の taxonomy と比較した際の利点、欠点を以下にそれぞれ挙げる。

3.2.1 利点

多面的な分類

従来の taxonomy ではあらかじめ専門家が構築した分類構造に基づいて、個々のコンテンツは厳密に一つの分類語に属さなければならない。それに対し、folksonomy では個々のユーザが自由にタグを付与し、個々のコンテンツはそれら複数の分類語に属するため多面的であるといえる。

低コスト

従来の taxonomy は、分類木構築フェーズと分類フェーズに分かれており、対象分野の専門家らが分類

木を構築する。それに対し、folksonomy では個々のユーザが自由にタグ付けを行うため専門家らが分類木を構築する必要がなく、低コストといえる。

分類が早い

folksonomy では何百～何万人ものユーザが分類を行うため分類速度が速い。特に新しいコンテンツに対する分類や新たな分類概念が発生した場合にも taxonomy よりも早く対応できる。

実用的な分類

folksonomy では各々のユーザが自分自身にとって最も使い勝手がよくなるようコンテンツにタグ付けを行うため、結果として多くのユーザにとって実用的な分類がなされる。厳密には間違った分類であったとしても、ユーザが認識している分類に従っているため、実用的であるといえる。

3.2.2 欠点

タグの表記ゆれ

folksonomy では、ユーザがタグに任意のキーワードを選択できるため、同じ意味の異なるタグが混在する事態が頻繁に起こる。これは Synonym (シノニム) と呼ばれ、例えば「ビジネス」、「business」、「BIZ」などが挙げられる。

タグの抽象度の違い

folksonomy では、個々のユーザが自分に使い勝手を基準にタグを選択するため、様々な抽象度のタグが混在してしまう。例えば、同じコンテンツに対してあるユーザが「サッカー」というタグを付与し、また別のユーザが「スポーツ」というタグを付与するといった状況が起こる。

タグの多義性

同じ表記で複数の意味を指す多義語タグが存在する。例えば、ソーシャルブックマークで使用される「app」というタグには「apple」という意味と「application」という意味の二つの意味がある。

ノイズ

自由にタグ付けするため、意味をなさないタグやスペル間違いのタグが存在する。例えば、ソーシャルブックマークでは意味のなさないタグとして「!!」などが見受けられる。

3.3 folksonomy と推薦システム

本節では前節で述べた folksonomy の利点、欠点が本論文の推薦システムにおいてそれぞれどのような効果をもたらすかについて述べる。まず利点の一つとして「多面的な分類」が本論文で目的とする嗜好に合致

し、かつまだ知らないコンテンツを推薦するのに有効であると考えている。一つのコンテンツに対し多面的にタグが付与されており、その中の一つの話題やテーマ単位で着目し推薦リストを多様化すれば、多様化に使われなかった話題やテーマにより推薦スコアは維持されるため、未知性の高い推薦が期待できると考えられる。

また既存の協調フィルタリングには cold-start 問題が挙げられるが [5], 「分類速度が速い」ことや「低コスト」であることから folksonomy のタグ情報を利用した推薦には、cold-start 問題の緩和が期待できる。タグ情報を使った推薦システムは、タグがコンテンツのメタ情報として機能するため、コンテンツベースと似た推薦結果になることが報告されている [18]。しかしながら、コンテンツ自体を解析する必要がないという利点もあり、特徴量が得られにくい種類のコンテンツにも対応可能である。本論文では、ソーシャルブックマークを対象としており、コンテンツの種類は Web ページだが、このような場合にも膨大な量の Web ページを解析し、特徴量を抽出する必要がないという点からタグ情報を用いた推薦手法が望ましいと考えられる。

以上より、folksonomy の利点が本論文で目的とする未知性の高い推薦に有用であるといえる。しかしながら、前述した欠点から推薦結果の予測精度を低下を招くため、本論文では階層的クラスタリング及びユーザプロファイルに基づいたタグクラスタの選択・粒度決定によりそれを解決する。

4. 提案手法

4.1 提案手法の概要

本論文では、ユーザのブックマーク情報に付与されたタグ群から嗜好情報を抽出し、それに基づき嗜好に合致し、かつ未知であるコンテンツを推薦することを目的とする。未知性を向上させるためには、ブックマーク数やタグ数の低いコンテンツや多様なコンテンツを推薦すればよいが、推薦されるコンテンツの質や推薦リストの予測精度（適合率）が低下してしまう可能性が高い。そこで提案手法では、あらかじめタグの表記ゆれを解消するためにクラスタリングをしておき、多面的に分類されたコンテンツに対し、タグクラスタ単位で推薦リストを多様化することにより、嗜好の範囲内だがまだ知らない内容のコンテンツを推薦する。推薦リストの多様化はタグクラスタに依存するため、多様化に用いるタグクラスタはユーザの嗜好に合致し、

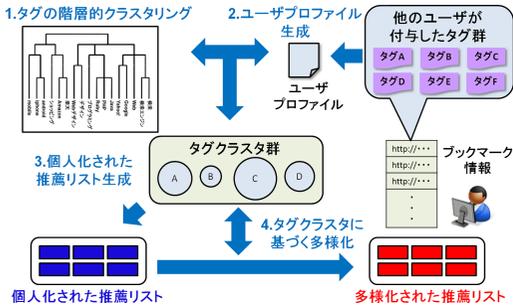


図 3 提案手法の概要

Fig. 3 Overview of out proposed method.

かつ嗜好の範囲内である必要がある。したがって提案手法では、ユーザのプロファイルに基づいてタグクラスタの選択・粒度決定を行い、そのタグクラスタに基づく多様化を行うことで予測精度を維持したまま未知性を向上させる。提案手法の概要を図 3 に示す。提案手法は以下のステップからなる。

- (1) タグの階層的クラスタリング
- (2) ユーザプロファイルの生成
- (3) 個人化された推薦リストの生成
- (4) タグクラスタに基づく推薦リスト多様化

ステップ 1 では、タグ頻度ベクトルを用いて、階層的クラスタリングを行う。ステップ 2 では、ユーザのブックマークから嗜好情報を抽出し、タグを用いたプロファイルを生成する。ステップ 3 では、ステップ 1, 2 の結果から個人化された暫定的な推薦リストを作成する。ステップ 4 では、ステップ 3 で推薦リストを生成するのに用いたタグクラスタを使って、個人化された推薦リストを多様化する。上記の方法によって、嗜好に合致し、かつまだ知らないコンテンツの推薦を、ユーザのブックマーク情報という少ない入力のみで実現する。以下の節では、各ステップにおける処理の詳細について述べる。

4.2 各ステップの詳細

4.2.1 タグの階層的クラスタリング

このステップでは、タグをベクトルとして表現し、それをもとにクラスタリングを行うことで、タグの表記ゆれの解消を試みる。folksonomy におけるタグの表記ゆれという問題に対してはクラスタリングが有効であるとされている [19]~[21], [23], [24]。またフォークソノミーでは、一つのタグが一つのクラスタとして生成される方が意味的なまとまりをより表現でき、最終的な検索や推薦の個人化の度合を高めることが

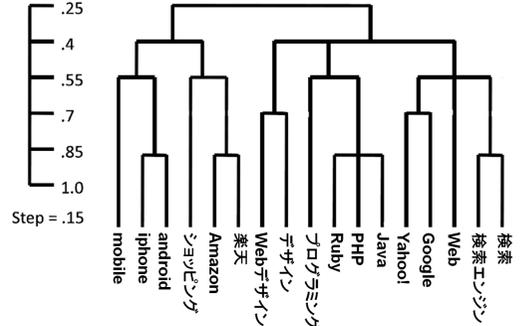


図 4 階層的クラスタリングの例

Fig. 4 An example of hierarchical clustering.

できるため、非階層的クラスタリングよりも階層的クラスタリングの方が効果的であると報告されている [21], [23], [24]。このことから、本論文においても、個々のユーザに対して、推薦リストの個人化を行っているため、このステップでは階層的クラスタリングが有効な手法であると考えられる。[21] の結果より、クラスタ間の距離については、重心法を用いる。なお本手法において、このステップはその後のステップの推薦処理とは独立しているため、ユーザは計算量の影響を受けない。以下にタグクラスタリングの流れを示す。

- (1) タグのベクトル化
- (2) 類似度算出
- (3) 類似度をもとにデンドログラム（樹形図）を構築

まずタグの各コンテンツに対するタグ頻度を属性とするベクトルを生成する。

$$\vec{t} = (w(c_1), w(c_2), \dots, w(c_n)) \quad (1)$$

式 (1) は、タグ t におけるベクトルを示しており、 $w(c_i)$ はタグ t のコンテンツ c_i におけるタグ頻度である。このベクトルをもとにタグ間の類似度を Cosine 類似度により算出する。この類似度に基づいてクラスタをマージしていく。全てのクラスタはサイズ 1 の個々のタグクラスタを作っている状態から始まり、クラスタ間の類似度がしきい値より大きければクラスタをマージしていく。クラスタがマージするかどうかを決定するしきい値はあらかじめ設定した Step 間隔で変化させ、全てのクラスタがマージされるまで繰り返すことにより、デンドログラム（樹形図）を構築する。図 4 は Step = 0.15 のときの階層的クラスタリングの例を示している。

4.2.2 ユーザプロファイルの生成

ユーザのブックマークを入力とし、そのブックマークされたコンテンツに付与されているタグ群からユーザプロファイルを算出し、ベクトルで表す。

$$\vec{u} = (w(u, t_1), w(u, t_2), \dots, w(u, t_n)) \quad (2)$$

式(2)は、ユーザ u におけるベクトルを示しており、 $w(u, t_i)$ はユーザ u とタグ t_i との関連度である。この関連度 $w(u, t_i)$ の算出方法を式(3)に示す。

$$w(u, t) = \sum_{c_i \in \text{bookmarks}} tfidf(t, c_i) \quad (3)$$

式(3)はユーザ u におけるタグ t の関連度を表しており、ユーザ u のブックマーク集合 bookmarks に含まれるコンテンツ c_i に付与されたタグの $tf\text{-idf}$ 値を加算していくことで算出する。 $tf\text{-idf}$ 値の算出方法を式(4)に示す。

$$tfidf(t, c) = \frac{tf(t, c)}{\sum_{t_i \in \text{tags}} tf(t_i, c)} \log \frac{\sum_{t_i \in \text{tags}, c_j \in \text{contents}} tf(t_i, c_j)}{\sum_{c_j \in \text{contents}} tf(t, c_j)} \quad (4)$$

式(4)において、 $tf(t, c)$ はコンテンツ c におけるタグ t のタグ頻度を示している。また tags はソーシャルブックマーク全体のタグ集合、 contents はソーシャルブックマーク全体の Web ページの集合をそれぞれ示している。 $tf\text{-idf}$ 値を算出することにより、「Java」や「PHP」といった特徴になりやすい具体的なタグの値を高く、「まとめ」や「あとで読む」といった抽象的なタグの値を低くすることができる。このユーザプロファイルをもとに個人化された推薦リストを生成する。

4.2.3 個人化された推薦リストの生成

提案手法では、多様化を行う前にあらかじめユーザごとに個人化された推薦リストを生成しておく。この個人化された推薦リストが最終的な推薦リストの推薦候補となる。ステップ1で構築したデンドログラム、ステップ2で生成したユーザプロファイルから個人化された推薦リストを生成する。処理の流れは以下のとおりである。

- (1) ユーザプロファイルに基づいたタグクラスタ

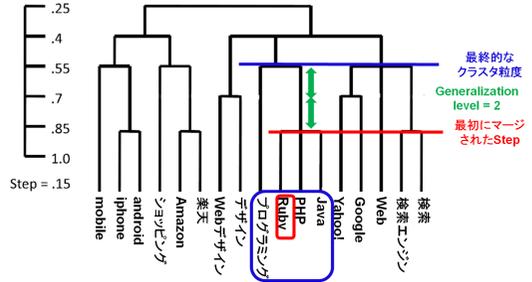


図5 タグクラスタの粒度決定
Fig. 5 Size determination of tag clusters.

の選択・粒度決定

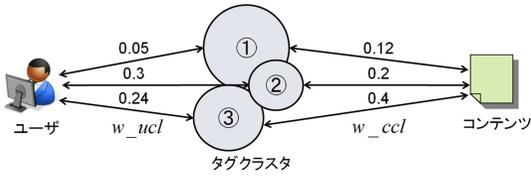
- (2) ユーザとタグクラスタの関連度算出
- (3) タグクラスタとコンテンツの関連度算出
- (4) 推薦スコア算出

本手法ではタグクラスタにより推薦リストを多様化するため、結果がタグクラスタに依存すると考えられる。したがって、予測精度と未知性の双方を高く保つためには、多様化に用いるタグクラスタはユーザの嗜好に合致する内容であり、かつ嗜好の範囲内である必要がある。そこで、個人化された推薦リストを生成するにあたって、まずユーザプロファイルに基づいてタグクラスタの選択・粒度決定を行う。ユーザプロファイルからしきい値 MinRelation 以上のタグが含まれるクラスタを選択し、そのタグがマージされたしきい値 Step から定数 $\text{GeneralizationLevel}$ 分粒度を大きくし、それを最終的なタグクラスタの粒度として決定する。例えばユーザプロファイルの「Ruby」というタグに着目した場合(図5)、「Ruby」を含むタグクラスタを選択し、粒度は「Ruby」タグがマージされたしきい値 $\text{Step} = 0.85$ よりあからじめ設定した定数 $\text{GeneralizationLevel} = 2$ という値に従って粒度を大きくする。そしてユーザプロファイルの関連度がしきい値 MinRelation 以上の全てのタグに対してこの処理を繰り返し、先に選択したタグクラスタと含有関係にある場合はクラスタサイズの大きい方のタグクラスタを選択する。この処理により、ユーザプロファイルで関連度の高いタグクラスタのみを複数選択できる。

推薦に用いるタグクラスタが決定したらクラスタとユーザ間の関連度を算出する。

$$w_ucl(u, cl) = \sum_{t_i \in cl} w(u, t_i) \quad (5)$$

式(5)はユーザ u とタグクラスタ cl の関連度を示して



$$recommend_score = 0.05 \times 0.12 + 0.3 \times 0.2 + 0.24 \times 0.4 = 0.162$$

図 6 個人化された推薦スコア算出例

Fig. 6 A calculation example of personalized recommendation score.

おり、選定したタグクラスタ cl に含まれるタグ t_i との関連度を加算していくことにより算出する。同様にしてタグクラスタとコンテンツの関連度を算出する。

$$w_ccl(c, cl) = \sum_{t_i \in cl} tfidf(t_i, c) \quad (6)$$

上記で算出したユーザ、タグクラスタ間の関連度とタグクラスタ、コンテンツ間の関連度をもとに個人化された推薦スコアを算出し、暫定的な推薦リストとする。

$$recommend_score(u, c) = \sum_{cl_i \in Clusters} w_ucl(u, cl_i) * w_ccl(c, cl_i) \quad (7)$$

式 (7) は推薦スコアを示しており、選択したタグクラスタ集合 $Clusters$ に含まれるタグクラスタ cl_i において、関連度の積を加算していくことにより算出する。図 6 に推薦スコアの算出例を示す。

4.2.4 推薦リストの多様化

このステップではステップ 3 で生成した暫定的な推薦リストをタグクラスタにより多様化する。ステップ 3 で生成した推薦リストにはある特定のタグクラスタの組合せにより似たようなコンテンツが多く含まれ、結果的に従来の協調フィルタリングやコンテンツベースと同様に既に知っているコンテンツが多く推薦されてしまうと考えられる。こうした問題を解決する方法の一つとして、推薦リストの多様化が有効であるとされている [2], [16], [17]。しかし、従来研究の分類方法で推薦リストを多様化した場合、コンテンツを一意に分類するため、ユーザが意図している分類に必ずしも一致するわけではない。したがって、推薦リスト中の嗜好に合致するコンテンツが減少してしまう可能性が高いと考えられる。そこで提案手法では、タグクラスタを用いて多様化することにより、ユーザの嗜好に合致しつつも未知なコンテンツを推薦する。具体的には、

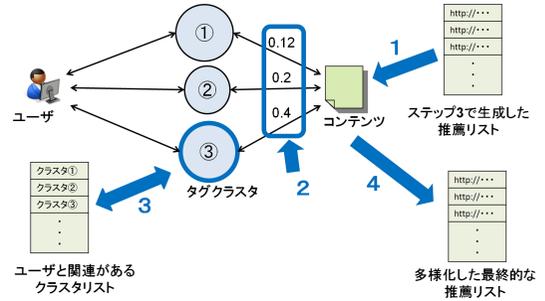


図 7 推薦リスト多様化アルゴリズムの例

Fig. 7 An example of recommendation list diversification algorithm.

タグクラスタと各コンテンツ間の関連度が最も高いタグクラスタに着目し、その関連度が最も高いタグクラスタが、最終的な推薦リストの N 件以内に 1 度しか現れないようにコンテンツを選択していくことにより、最終的な推薦リスト N 件を多様化する。推薦リスト多様化のアルゴリズムの流れを以下に示す。

- (1) 個人化された推薦リストの上位のコンテンツから順に選定する
- (2) 取り出したコンテンツと最も関連度の高いタグクラスタを調べる
- (3) そのタグクラスタがユーザと関連度のあるクラスタの中でまだ選ばれていなければ、多様化した最終的なリストに追加する
- (4) 多様化した最終的な推薦リストがあらかじめ設定した N 件になるまで繰り返す

図 7 は推薦リストを多様化する際の例を示している。まずステップ 3 で生成した推薦リストの上位から順の一つずつコンテンツを選定していく (図 7 中の 1 に該当)。その選定したコンテンツと最も関連度の高いタグクラスタを調べる (図 7 中の 2 に該当)。図 7 の例では、上から三つ目のタグクラスタが最も関連度が高い。そして、そのタグクラスタがユーザと関連するタグクラスタの中でまだ選ばれていなければ (図 7 中の 3 に該当)、新たな多様化された推薦リストにコンテンツを追加し (図 7 中の 4 に該当)、あらかじめ設定した N 件に達するまで繰り返す。こうした処理をすることにより、追加したコンテンツはある程度多様化されるが、ユーザと複数のタグクラスタを介して関連しているので高いスコアが維持でき、嗜好に合致しつつもまだ知らない可能性の高いコンテンツが推薦されやすくなると考えられる。

Algorithm 1 推薦リスト多様化アルゴリズム.

Input: $List_i$: 個人化された推薦リスト, $TagClusters$: ユーザと関連度があるクラスタ集合
Output: $List_i^D$: 多様化された推薦リスト

```

1: for all content in  $List_i$  do
2:    $MostRelTagCluster$ 
   ← { $GetMostRelTagCluster(content)$ };
3:   if has( $TagClusters, MostRelTagCluster$ ) then
4:     add( $List_i^D, content$ );
5:     delete( $TagClusters, MostRelTagCluster$ );
6:   end if
7:   if  $N = |List_i^D|$  then
8:     return  $List_i^D$ ;
9:   end if
10: end for

```

詳細なアルゴリズムに関しては Algorithm 1 に示す。ステップ 3 で生成した個人化された推薦リスト $List_i$ と対象ユーザと関連度があるクラスタ集合 $TagClusters$ を入力とし、最終的な多様化された推薦リスト $List_i^D$ を出力する。 $List_i$ はスコアに基づいて降順にコンテンツが格納されているものとする。まず最初の for 内で、 $List_i$ から順にコンテンツを取り出す。そして、そのコンテンツと対象ユーザ双方に関連しているタグクラスタのうち最も関連度の高いタグクラスタを取り出す ($GetMostRelTag$)。そのタグクラスタがまだ選ばれていないタグクラスタであれば ($has(TagClusters, MostRelTagCluster)$)、 $List_i^D$ に追加し ($add(List_i^D, content)$)、あらかじめ設定した N 件に達するまで繰り返す ($N = |List_i^D|$)。

5. 評価実験

5.1 実験の概要

5.1.1 データセット

本論文で提案する手法がソーシャルブックマークにおける推薦リストの未知性向上に有効であるかを検証するため、アンケートによる推薦リストの評価実験を実施した。実験データとして、はてなブックマークにおける 2009 年 1 月～2010 年 12 月までのデータセットから、タグ付け回数 100 回以上のユーザをランダムに 5000 人選び、利用した。そこから前処理を行い、全体でタグ頻度 10 回以下のタグをノイズとして削除した結果、データサイズとしてコンテンツ (Web ページ) 数 115,944, タグ数 10,599, ユーザ数 5000 人で、その組み合わせであるブックマークデータ 3,591,302 レコードの実データを用い、実験を行った。

表 1 評価実験時のパラメータの値

Table 1 Parameter values of our experiment.

MinRelation	1.183
Step	0.1
GeneralizationLevel	2

5.1.2 実験手順

10 名の被験者により実験を行った。ソーシャルブックマーク自体は利用していないが、日常的にブラウジングでブックマークを利用しており、ブックマーク数が 20～60 件の被験者をそろえた。被験者には普段利用しているブックマークを提供してもらい、そのブックマークを入力として推薦リストを作成し、提案手法、タグベース手法、タグクラスタ手法及びロングテール手法のそれぞれの推薦リストに対するアンケート調査を実施した。

推薦された上位 10 件の Web ページに対して嗜好に合致し、かつ未知であるかを評価してもらうため、興味について (1:非常に興味がある, 2:どちらかといえば興味がある, 3:どちらかといえば興味がない, 4:全く興味がない) の 4 通りの評価と興味があるページのうち知っている内容であるかどうか (1:知っている, 2:知らない) の 2 通りで、Web ページの内容を閲覧し、評価してもらった。また、本論文では folksonomy のタグ情報のみを利用し、ユーザの嗜好内で嗜好に合致し、かつ未知なコンテンツを推薦することを目的としているので、ベースラインとしてタグ頻度ベクトルの類似度により似たものを推薦するタグベース手法、ステップ 3 で生成した推薦リストであるタグクラスタ手法、及び代表的な手法として Wu ら [10] の手法を元にしたロングテール手法の三つと比較した。タグベース手法は、コンテンツにタグが付与された回数を属性値としたコンテンツベクトルと、ユーザのブックマークしたコンテンツに付与されたタグの回数を集計し、属性値としたブックマークベクトルとの Cosine 類似度が高い順に推薦する手法である。また評価してもらう際、手法間の不公平さをなくすため、それぞれ上位 10 件ずつの Web ページをランダムに混ぜ、計 40 件を評価してもらった。なお推薦リストのコンテンツが重複した場合には、同じ評価値をつけてもらった。提案手法のパラメータは先行研究 [23] を参考に、表 1 のように設定した。

5.2 実験結果

被験者が付けた興味に関する評価の平均値を求めた。

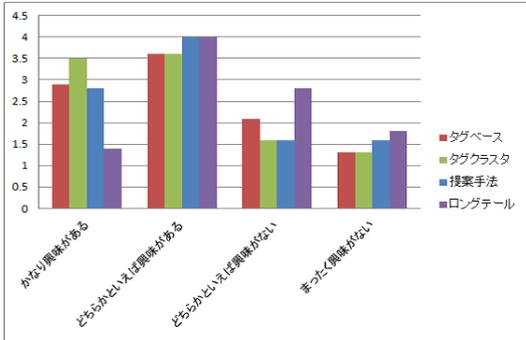


図 8 推薦リストに対する興味があるかの評価
Fig. 8 Rating of interest in the recommendation list.

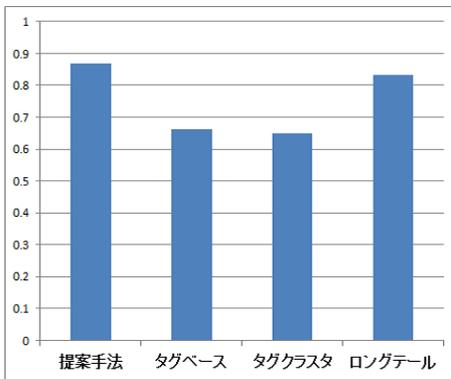


図 9 各推薦リスト 10 件中の「興味がある」と「興味があるかつ未知である」コンテンツの割合
Fig. 9 Rate of interest and unknown contents in the recommendation list.

結果を図 8 に示す。図 8 より、「かなり興味がある」の割合はタグクラスタ手法が最も高く、次いでタグベース手法が高いことが確認できる。タグクラスタ手法と提案手法の間には有意な差が確認されたが、提案手法とタグベースの間には有意な差はみられなかった。この結果より提案手法はタグベースと同等の予測精度（適合率）であることが分かる。また、提案手法とロングテール手法、タグベース手法とロングテール手法及びタグクラスタ手法とロングテール手法の間にはそれぞれ 5% 有意水準にて有意な差が確認できた。この結果より、ロングテール手法では未知性と引き換えに興味あるコンテンツを推薦できる可能性が減ってしまうということがいえる。

興味がある（1:非常に興味がある, 2:どちらかといえば興味がある）コンテンツのうち知らないコンテンツの割合の平均値を求めた。結果を図 9 に示す。図 9

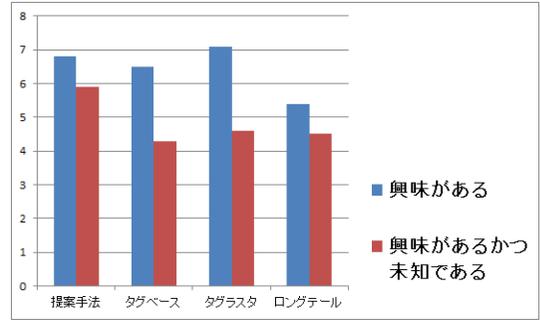


図 10 各推薦リスト 10 件中の「興味がある」と「興味があるかつ未知である」コンテンツの平均個数
Fig. 10 Average of content which is "interest" and "interest and interested" in the recommendation list.

より、推薦結果の興味があるコンテンツのうち知らないコンテンツの割合はロングテール手法と同程度の値である。また提案手法とタグベース手法、ロングテール手法とタグベース手法、提案手法とタグクラスタ手法及びロングテール手法とタグクラスタ手法の間にはそれぞれ 5% 有意水準にて有意な差が確認できた。この結果より提案手法はロングテール手法と同程度未知性の高い推薦が実現できているといえる。

被験者が付けた評価のうち「非常に興味がある」「興味がある」と「興味があるかつ知らない」の平均値を求めた。結果を図 10 に示す。図 10 より、提案手法は「興味がある」かつ未知なコンテンツの割合が最も高いことが確認できる。また、興味があるかつ未知なコンテンツにおいて提案手法とタグベース手法、提案手法とタグクラスタ手法及び提案手法とロングテール手法の間にはそれぞれ 5% 有意水準にて有意な差が確認できた。したがって、これらの結果より、タグベース手法、タグクラスタ手法及びロングテール手法に比べ、提案手法では「興味があるかつ未知である」コンテンツを多く推薦できているといえる。

6. 考 察

本章では推薦リストの予測精度（適合率）、未知性、提案手法の有効性について考察する。

実験結果より、提案手法が従来のタグベース手法、タグクラスタ手法及びロングテール手法の推薦リストよりも嗜好に合致し、かつまだ知らないコンテンツを多く含んでいることが確かめられた。本論文では、推薦リストの未知性の向上のため、ユーザの嗜好に合致するかつまだ知らない内容であるコンテンツを推薦す

ることを目的としている。ユーザがまだ知らない内容のコンテンツを推薦することと嗜好に合致するコンテンツを推薦することの間には、一般的にトレードオフの関係があるといわれているが、評価実験では提案手法の予測精度（適合率）として約 60%~70%、興味があるかつ未知なコンテンツの割合として約 60%という値がそれぞれ導き出された。タグ情報のみを用いた推薦システムの予測精度としては、既存手法 [19], [20] が約 40%~60%であることから比較的良好な結果であるといえる。またタグに限らず一般的な範囲を対象とした推薦システムとしても、遜色ない値であるといえる。未知性に関しては、図 9 より推薦結果の興味があるコンテンツのうちまだ知らないコンテンツの割合はロングテール手法と同程度の値であるが、図 8 より提案手法は適合率を維持できているため、興味があるかつ知らないコンテンツを推薦できる可能性は高い。したがって、本論文の提案手法では、folksonomy の欠点から生じる予測精度の低下をクラスタリング及びタグクラスタの選択・粒度決定で解決し、そのタグクラスタに基づいて多様化することにより、適合率を維持しつつも、未知性の高い推薦が実現できているといえる。

提案手法では、ユーザと複数のタグクラスタを介して推薦されたコンテンツに対し、タグクラスタ単位で推薦リストを多様化している。そのため、その複数のタグクラスタがユーザにとって興味がある内容であり、かつ興味の範囲内で選択されている必要がある。タグクラスタの粒度が大きすぎると推薦リストの多様化が荒く、未知なコンテンツは増えるが予測精度は下がってしまう。逆に粒度が小さすぎると予測精度を維持できるが、多様化が不十分で未知なコンテンツが減少してしまう。提案手法では、ユーザプロフィールに基づいて、対象分野ごとにタグクラスタを選択・粒度決定を行い、推薦に利用するタグクラスタを個人化している。こうした処理が、予測精度低下の原因となっているタグの表記ゆれや興味の範囲外に属するタグクラスタの問題を解決しており、結果的に適合率の維持につながったと考えられる。また未知性が向上していることに関しては、この個人化したタグクラスタに着目して推薦リストを多様化したため、ユーザの嗜好内で嗜好傾向の幅が広くなり、結果的に既知なコンテンツが推薦される可能性が減ったことやユーザが今まで気付かなかったタグクラスタの組合せにより新たなコンテンツが推薦されたことが要因と考えられる。

しかし提案手法では、ユーザプロフィールに基づい

て選択したタグクラスタがコンテンツに対して多面的に関連していない（一つのタグクラスタのみを介す）場合、タグの多義性による問題が顕著にでてしまうため、意図しないコンテンツが推薦されてしまう傾向にある。提案手法ではしきい値を一律に設定しているが、入力ブックマーク数やそのブックマークに付与されたタグにどれほど多様性があるかなどを考慮してパラメータを設定することで、解決できると考えられる。

7. む す び

本論文では、folksonomy 型サービスにおける推薦の未知性向上を目的として、嗜好に合致するが知らないコンテンツを推薦する手法を提案した。提案手法では、folksonomy 特有の多面的に付与されたタグに基づいて推薦リストを多様化することにより、推薦リスト中の嗜好に合致し、かつ知らないコンテンツが増加するという仮説をもとに、未知性を考慮した推薦手法を提案した。またその際 folksonomy の欠点から生じる予測精度の低下に対し、階層的クラスタリング及びユーザプロフィールに基づくタグクラスタ選択・粒度決定による解決を試みた。評価実験では、実際にソーシャルブックマークサービスはてなブックマークのデータを用いて、普段被験者が利用しているブックマーク情報を入力にコンテンツを推薦し、アンケートによる評価を行った。その結果、予測精度（適合率）に関しては約 60~70%と高い値を維持することができ、またユーザの嗜好に合致するが、知らないコンテンツについても既存のタグベース手法、ロングテール手法より多く推薦できることが確認できた。また提案手法において、ユーザが今まで気付かなかったタグクラスタの新たな組合せにより、知らないコンテンツが推薦される傾向にあり、本手法では考慮していなかった推薦の意外性という観点に関しての発展性も期待できると考えられる。

今後の課題として、今回はブックマーク件数が 20~60 件のユーザに適用したが、より広範囲のユーザに対しての有効性を検証していきたい。また最適なパラメータを設定するために、入力ブックマークの多様性と推薦に用いるタグクラスタ数や出力との関連性について分析していきたいと考えている。

最後に、本論文では folksonomy 型のサービスであるソーシャルブックマークを対象にしたが、これに限らず、以下のような要件を満たす folksonomy 型サービスに対し本手法は適用可能である。

- タグが付与できる
- ユーザが気に入ったコンテンツをブックマークできる

特にタグの表記ゆれが頻繁に見受けられ、コンテンツが複数分野のタグで多面的に分類されている folksonomy 型サービスにおいて、本手法は有効であると考えられる。

謝辞 本研究を遂行するにあたり、研究の機会と議論・研鑽の場を提供して頂き、御指導頂いた国立情報学研究所/東京大学本位田真一教授をはじめ、活発な議論と貴重な御意見を頂いた研究グループの皆様へ感謝致します。

文 献

- [1] J.B. Schafer, J.A. Konstan, J. Riedl, "E-commerce recommendation applications," *Data Mining and Knowledge Discovery*, vol.5, pp.115-153, 2001.
- [2] C.N. Ziegler, S.M. Mcnee, J.A. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," *World Wide Web Conference*, pp.22-32, 2005.
- [3] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Trans. Information Systems*, vol.22, no.1, pp.5-53, 2004.
- [4] S.M. McNee and J. Riedl, "Being accurate is not always good: How accuracy metrics have hurt recommender systems," *ACM Special Interest Group on Computer Human Interaction*, pp.1103-1108, 2006.
- [5] 神畷敏弘, "推薦システムのアルゴリズム (1)," *人工知能誌*, vol.22, no.6, pp.826-837, 2007.
- [6] Delicious, <http://delicious.com/>
- [7] はてなブックマーク, <http://b.hatena.ne.jp/>
- [8] flickr, <http://www.flickr.com/>
- [9] Youtube, <http://www.youtube.com/>
- [10] X. Wu, Y. Zhang, J. Guo, and J. Li, "Web video recommendation and long tail discovering," *2008 IEEE International Conference on Multimedia and Expo*, pp.369-372, 2008.
- [11] 住元宗一郎, 中川博之, 田原康之, 大須賀昭彦, "コンテンツ投稿型 SNS における未知性と意外性を考慮した推薦エージェントの提案," *信学論 (D)*, vol.J94-D, no.11, pp.1800-1811, Nov. 2011.
- [12] Y. Park and A. Tuzhilin, "The long tail of recommender systems and how to leverage it," *ACM Conference on Recommend System*. 2008, 2008.
- [13] Y. Hijikata, T. Shimizu, and S. Nishida, "Discovery-oriented collaborative filtering for improving user satisfaction," *ACM Conference on Intelligent User Interfaces 2009*, pp.67-76, 2009.
- [14] 清水拓也, 土方嘉徳, 西田正吾, "発見性を考慮した協調フィルタリングアルゴリズム," *信学論 (D)*, vol.J91-D, no.3, pp.538-550, March 2008.
- [15] Pixiv, <http://www.pixiv.net/>
- [16] C.N. Ziegler, G. Lausen, and L.S. Thieme, "Taxonomy driven computation of product recommendations," *2004 ACM CIKM Conference on Information and Knowledge Management*, pp.406-415, 2004.
- [17] 小川祐樹, 諏訪博彦, 山本仁志, 岡田 勇, 太田敏澄, "動的なトピック分類に基づく Novelty を考慮した推薦アルゴリズムの提案," *情処学論*, vol.50, no.6, pp.1626-1648, 2009.
- [18] I. Guy, N. Zwerdling, I. Ronen, D. Carmel, and E. Uziel, "Social media recommendation based on people and tags," *ACM Special Interest Group on Information Retrieval*, pp.194-201, 2010.
- [19] S. Niwa, T. Doi, and S. Honiden, "Web page recommender system based on folksonomy mining," *3th International Conference on Information Technology, New Generations*, 2006.
- [20] 丹羽智史, 土肥拓生, 本位田真一, "Folksonomy マイニングに基づく Web ページ推薦システム," *情処学論*, vol.47, no.5, pp.1382-1392, 2006.
- [21] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke, "Personalized recommendation in social tagging systems using hierarchical clustering," *ACM Conference on Recommend System*, 2008.
- [22] A. Mathes, "Folksonomies-cooperative classification and communication through shared metadata," *Computer Mediated Communication*, vol.47, no.10, 2004.
- [23] J. Gemmell, A. Shepitsen, M. Mobasher, and R. Burke, "Personalization in folksonomies based on tag clustering," *Proc. 6th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems*, 2008.
- [24] J. Gemmell, A. Shepitsen, B. Mobasher, and R. Burke, "Personalizing navigation in folksonomies using hierarchical tag clustering," *Proc. 10th International Conference on Data Warehousing and Knowledge Discovery*, 2008.

(平成 24 年 6 月 5 日受付, 10 月 3 日再受付)

藤原 誠



2012 電気通信大学大学院情報システム学研究科社会知能情報学専攻修士課程了, 同年 (株)野村総合研究所入社, 現在に至る。



中川 博之 (正員)

1997 阪大・基礎工・情報工学卒。同年鹿島建設(株)に入社。2007 東京大学大学院情報理工学系研究科修士課程了。2008 同大学院博士課程中退。同年より電気通信大学助教。現在に至る。エージェント及び自己適応システム開発手法の研究に従事。情報処理学会, IEEE CS 各会員。



田原 康之

1991 東京大学大学院理学系研究科数学専攻修士課程了。同年(株)東芝入社。1993~1996 情報処理振興事業協会に外向。1996~1997 英国 City 大学客員研究員。1997~1998 英国 Imperial College 客員研究員。2003 国立情報学研究所入所。2008 より電気通信大学准教授。博士(情報科学)(早稲田大学)。エージェント技術, 及びソフトウェア工学などの研究に従事。情報処理学会, 日本ソフトウェア科学会会員。



大須賀昭彦 (正員)

1981 上智大・理工・数学卒。同年(株)東芝入社。同社研究開発センター, ソフトウェア技術センターなどに所属。1985~1989 (財)新世代コンピュータ技術開発機構(ICOT) 外向。2007 より, 電気通信大学大学院情報システム学研究科教授。工博(早稲田大学)。主としてソフトウェアのためのフォーマルメソッド, エージェント技術の研究に従事。1986 年度情報処理学会論文賞受賞。情報処理学会, 日本ソフトウェア科学会, 人工知能学会, IEEE CS 各会員。