放送型情報配信システムのための時系列性を考慮した情報フィルタリング

馬 強[†] 角谷和俊[††] 田中克己[†]

近年，インターネットにおける放送型ニュース配信システムが注目を集めている．放送型ニュース配信システムでは，配信される記事は時間的に追加もしくは更新される，いわば，時系列データである．ニュース記事の情報価値は時間に関連しているので，ユーザにとって価値の高い情報を選択するには，従来の手法は不十分である場合がある．本論文では，フィルタリングなどの情報処理に必要となる配信情報の特徴量（新鮮度・流行度・緊急度）を，配信記事のコンテンツと配信時間および配信履歴に基づいて定義し，これに基づいて，ユーザプロファイルと併用した情報フィルタリング手法を提案する．

Information Filtering Based on Time-series Features for Data Dissemination Systems

Qiang Ma,[†] Kazutoshi Sumiya[††] and Katsumi Tanaka[†]

Broadcasting-type information dissemination systems on the Internet are becoming increasingly popular due to advances in the area of Web technology and information delivery. One of the notable features of push-based, multiple-channel-based information dissemination systems is to send information to users in a form of time-series articles. Conventional information filtering method does not consider well the *worth* of an article from the standpoint of the *time-series feature*. In this paper, the worth of an article compared with past delivered articles is considered with its time-series features (freshness, popularity, urgency), based on both its contents and temporality. Based on both of the user profile and the time-series features, we propose a new information filtering method and show some experimented results.

1. はじめに

近年，インターネットにおける情報配信システム[1)]が注目を集めている．特に，WWW（World Wide Web）による情報提供サービスが急激に増加し，さまざまな情報を共有することが可能になってきている．しかし，提供される情報が多くなるにともなって，ユーザは自分が欲しい情報を探し出すことが困難になっている．

このような問題を解決するために，欲しい情報をユーザが能動的に探し出すのではなく，情報を自動的に配信する放送型情報配信システムが提案されている[2)~10)]．これらのシステムでは，ユーザによってあらかじめ設定されたプロファイルに従って情報フィルタリングを行う．ユーザが受動的に情報を受信できるので，大量の情報にアクセスできるとともに，情報を獲得するためのユーザの負担が減少される利点がある．

放送型情報配信システムでは，配信される情報は時間とともに変化する．つまり，時系列データである．たとえば，時系列的に配信されるニュース記事は，次のような特性を持つ：

- 内容の未知性：ニュースの内容は幅広く，新しい情報であるため，予測不可能である．つまり，内容は未知である．
- 時系列性：ニュースは日々更新・追加され，連続的に配信される．
- 関連性：ほとんどのニュース記事は，続報や前報☆など関連記事が存在する．これらの記事には関連性がある．

大量の情報からユーザがほしい情報を見つけるため

† 神戸大学大学院自然科学研究科
Graduate School of Science and Technology, Kobe University
†† 神戸大学都市安全研究センター
Research Center for Urban Safety and Security, Kobe University

☆ 記事 a_0 の続報記事が a_1 である場合，a_0 を a_1 の前報と呼び，以下前報と記す．

に，ユーザプロファイルを用いた情報フィルタリングがよく利用される．ユーザプロファイルによるフィルタリングは，あらかじめユーザが指定したキーワードに基づいて，ユーザプロファイルを作成してフィルタリングを行い，フィルタされて得られた結果に対して，ユーザが妥当性をフィードバックし，ユーザプロファイルを修正していく手法である．しかし，これらの手法はニュース記事の上記のような特性を配慮していないため，キーワードであらかじめ指定できないような未知のニュースは獲得困難となる場合があり，放送型ニュース配信システムのフィルタリング機構にさらなる工夫が必要であると考えられる．

- ニュースの内容は未知であるので，適切なユーザプロファイルの定義が困難である場合がある．
- ニュース記事の間に関連性が高いので，前報や続報などの関連記事の記事の価値への影響を配慮する必要がある．

本論文では，ニュース記事の内容未知性，時系列性と関連性という特性を考慮して，フィルタリングを行うときに必要となるニュース記事の特徴量を，配信記事のコンテンツと配信時間および過去の配信履歴など時間関連要素に基づいて定義し，ユーザプロファイルと併用したフィルタリング手法を提案する．

以下，本論文の構成を示す．2 章では，関連研究について述べる．3 章では，放送型情報配信システムにおける記事の時系列的特徴量について述べる．特に，過去の配信履歴に依存する新鮮度と流行度，更新頻度に基づく緊急度の概念を導入する．そして，これらの特徴量と従来のユーザプロファイルを併用したフィルタリング手法を提案する．4 章では，ニュース記事の時系列的特徴量とユーザプロファイルを併用したフィルタリング方式と，従来型のフィルタリング方式を実験を通じて比較し，時系列的特徴量を用いたフィルタリングの効果を評価する．また，新鮮度と遡及範囲（新鮮度を計算するために遡った過去の配信履歴の範囲）の関係を実験結果を用いて考察を行う．5 章では，まとめと今後の課題について述べる．

2. 関連研究

Pointcast [11),12)] は，最も普及している放送型（プッシュ型）情報配信システムの 1 つである．Pointcast は，ユーザのキーワード・プロファイルなどによる情報フィルタリングは行えず，ユーザの登録したチャンネルのニュース記事を配信する．ユーザによる選択はチャンネルの追加・削除のみに限られ，ユーザ個人の要求を十分に満たせない場合がある．

スタンフォード大学では，publish/subscribe モデルを用いた情報配信システム SIFT [13)] が開発されている．SIFT では，明示的なユーザプロファイルの定義とユーザのフィードバックを用いて情報フィルタリングを行えるので，ユーザの多様な要求を満たすことができるが，本論文で提案する時系列的特徴量によるフィルタリングは考慮されていないので，情報の重複や新鮮度の高い記事の漏れなどの問題が起こる可能性がある．

Shapiro は，大量の Web コンテンツを動的にフィルタリングを行うことによってチャンネルを生成する手法を提案している [14)]．Shapiro は，プル型のユーザインタフェースをプッシュ型に移行することを目標としているが，チャンネルを流れる情報は Web コンテンツを対象としており，本論文で取り上げているような配信コンテンツの時系列性を考慮したフィルタリングは考慮していない．

サイト・アウトライニング [15),16)] では，Web サイトを動的な情報源としてとらえ，情報検索を支援するためのメタ情報の抽出手法を提案している．サイト・アウトライニングでは，情報の掲載日時，掲載期間と掲載回数など時系列変化に注目しているが，本論文のように，過去の配信記事との非類似性に基づく記事の新鮮度計算などは考慮していない．

ANATAGONOMY [17)] はユーザアクセスを監視し，適合フィードバックすることで，ユーザが興味のある情報を自動的に呈示するプッシュ型の情報配信システムである．ニュース記事のようなコンテンツの時系列性を考慮していない点が本研究と異なる．

Massachusetts 大学の Allan ら [18)]，Carnegie Mellon 大学の Yang ら [19)] は TDT（Topic Detection and Tracking）に関する研究を行っている．彼らは，既存の記事またはオンラインニュースから新しい話題（topic）の検知と追跡を行う手法について研究を行っている．彼らは記事のクラスタリングやクエリーなどの手法を用いて，話題の切り分けを行っているが，本論文では，新しく配信された記事を，過去の配信記事との類似・非類似性を計算してその記事の新鮮度・流行度を評価してフィルタリングする点が異なっている．

宗像ら [20)] は，周期的に発生するデータ系列から，データの鮮度と同期度に基づいてデータの組合せを選択する手法を提案している．彼らは，データの鮮度を，得られたデータの中で一番古いデータの時刻から現在の時刻までの経過時間として定義している．データの同期度は，得られたデータの中で，一番新しいデータの時刻と一番古いデータの時刻の差として定義して

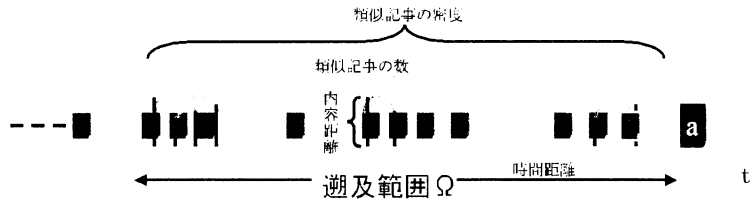

図 1 新鮮度
Fig. 1 Freshness.

いる．宗像らの鮮度は，固定的な周期ごとに発生するデータを対象としているが，ニュースの場合，記事は周期的に更新されるとは限らない．

3. ユーザプロファイルと時系列的特徴量に基づくフィルタリング

放送型ニュース配信では，前報や続報といった関連記事があるので，ユーザにとって価値の高い情報を選択するためには，従来のキーワードのみの情報フィルタリング手法では不十分であると考えられる．本章では，配信記事のコンテンツと配信時間および過去の配信履歴に基づいて記事の時系列的特徴量を定義し，これらの時系列的特徴量とユーザプロファイルを併用したフィルタリング手法を提案する．

3.1 時系列的特徴量計算のための遡及範囲

ニュース記事は関連する記事などが時系列的に配信されるので，各ニュース記事の価値は過去に配信された関連記事に依存する．

一般に，文書の特徴量を表す特徴ベクトルは tf・idf 法[21] を用いて計算されることが多いが，idf 値は対象となる文書集合に依存する．特に，放送型ニュース配信システムの場合，各記事の tf・idf 値を計算する場合，過去に配信された記事をどの程度まで遡って idf 値の計算の対象とするかが重要である．

本論文では，以下のように，時間幅と記事数の 2 つの基準に基づいて idf 値計算のための遡及範囲を選択する．

- 時間幅　フィルタリングの結果として得られた過去の配信記事のうち，ある時間範囲にある記事の集合を idf 値計算のための文書集合とする方法である．たとえば，現在時間から 1 週間以内の記事で，フィルタリングの結果得られたものを処理集合とする．この場合，記事数は不定である．
- 記事数　フィルタリングの結果として得られた過去の配信記事のうち，指定された数の記事を持つ記事の集合を遡及範囲とする方法である．記事数の場合，時間幅は不定となる．

3.2 時系列的特徴量

時間の経過とともに配信されるニュース記事の価値は，その記事自身の内容だけでなく，過去の配信履歴，配信頻度など要素にも依存する．本論文では，ニュース記事の時系列的特徴量として，配信履歴に基づく新鮮度と流行度，および配信頻度に基づく緊急度を定義する．

3.2.1 新 鮮 度

新鮮度の計算とは，一定の時空間において，オブジェクト（記事）を既存のオブジェクト集合と比べて，そのオブジェクトの新しさに対する評価を行うことである．

図 1 に示すように，ある遡及範囲（記事の集合）Ω に対する記事 a の新鮮度を，(1) Ω 内の類似記事の数に基づく新鮮度（$fresh_{num}(a)$），(2) Ω 内の類似記事との内容距離に基づく新鮮度（$fresh_{cd}(a,\omega)$），(3) 記事の密度に基づく新鮮度（$fresh_{de}(a)$），(4) 類似記事との時間距離に基づく新鮮度（$fresh_{td}(a,\omega)$）によって定義する．これらの新鮮度は，ユーザの選択によって，独立に用いることができる．次に，これらの新鮮度を混合して，それぞれに重みを付けた統合新鮮度 $fresh_{\Omega}(a)$ を次式のように定義する．

$$\begin{aligned} fresh_{\Omega}(a) = {} & \alpha * fresh_{num}(a) && (1)\\ & + \beta * fresh_{cd}(a,\omega) && (2)\\ & + \gamma * fresh_{de}(a) && (3)\\ & + \sigma * fresh_{td}(a,\omega) && (4) \end{aligned}$$

ただし，ω は Ω における a の類似記事の集合である．α，β，γ，σ は重み付け定数である．

ここで，記事 a の新鮮度を計算するための遡及範囲 Ω の記事数を n とする．Ω における a の類似記事の集合 ω の記事数を m とする．

(1) 類似記事の数による新鮮度

過去の配信記事の集合 Ω の中で，a と類似している記事の数が少なければ，その記事の内容が新しく，新鮮度が高いと考えられる．そこで，類似記事数に基づく新鮮度を次のようにする．

$$fresh_{num}(a) = \frac{1}{\log_2(2+m)} \quad (5)$$

(2) 内容距離による新鮮度

各記事 a は，k 次元特徴ベクトル $v(a) = (w_1, w_2, \ldots, w_k)$ を持つとする．ただし，w_i はキーワード k_i の重みである．

記事 a と b の違いを表す内容距離 $dis(a,b)$ を次のように定義する：

$$dis(a,b) = |v(a) - v(b)| \quad (6)$$

つまり，$v(a)$，$v(b)$ の差を記事 a，b の内容距離とする．

記事 a が前報記事（過去の類似記事）と比べてどの程度新しい情報を追加しているかを，記事 a と前報記事との内容距離で表すことができる．前報の記事とは異なる情報が多く追加される場合（内容距離が大きい場合），a の新鮮度が高いと考えられる．つまり，記事 a の類似記事集合 ω ($\omega \subseteq \Omega$) 内の記事との平均内容距離が大きいほど a の新鮮度が高いと考えられる．よって，記事 a の類似記事集合 ω に対する新鮮度を

$$fresh_{cd}(a,\omega) = \log\left(\frac{1}{m}\sum_{i=1,b_i\in\omega}^{m} dis(a,b_i)\right) \quad (7)$$

と定義する．

(3) 類似記事の密度による新鮮度

a の類似記事の Ω における密度は $d = m/n$ である．d を a の予想出現確率と考えると，a の情報量は $\log_2 \frac{1}{d}$ となる．情報量が大きいと，新鮮度が高いと考えられるので，記事 a の類似記事密度に基づく新鮮度を

$$fresh_{de}(a) = \log_2 \frac{n}{m} \quad (8)$$

と定義する．

(4) 時間距離による新鮮度

過去の類似記事との時間距離も記事 a の新鮮度に影響を与える．類似記事と時間的に離れれば離れるほど a の新鮮度が高くなると考えられる．直観的には，たとえば，**図 2** では，$(a.1)$ の a の新鮮度は $(a.2)$ の a より大きい，$(b.2)$ の a の新鮮度は $(b.1)$ の a より小さい．

a と類似記事集合 ω の平均時間距離が大きければ大きいほど a の新鮮度が大きくなると考え，記事 a の類似記事集合 ω との時間距離に基づく新鮮度を次式のように定義する．

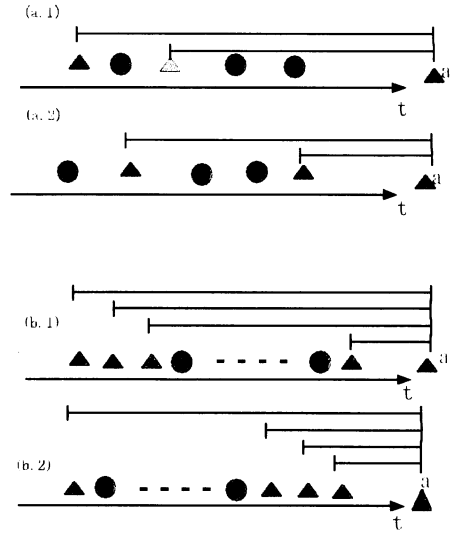

図 2 時間距離の例
Fig. 2 Example of time distance.

$$fresh_{td}(a,\omega) = \log\left(\frac{1}{m}\sum_{i=1,b_i\in\omega}^{m}(t(a) - t(b_i))\right) \quad (9)$$

ただし，$t(a)$ は記事 a の配信時間を表す．

3.2.2 流 行 度

最近の配信記事の中で，新しい配信記事と類似しているものが多数存在する場合，この記事はあるトピックの続報であり，現在話題になっていると考えられる．このような記事は，新鮮度が低いが，流行度が高く，ニュース価値が高いと考えられる．新鮮度と流行度の計算では，ともに過去の配信記事との比較を行っているので，流行度と新鮮度は基本的にお互いに依存する．ただ，遡及範囲の選択方式などが異なると依存度が変化する場合があり，本論文では，新鮮度と流行度を別個のものとして扱っている．

本論文では，記事 a の流行度 $pop(a)$ は類似記事の密度と時間距離を用いて定義する（**図 3**）．つまり，最近の配信記事の中で，a との類似記事が多ければ多いほど，a の流行度が高くなると考え，次式のように定義する．

$$pop(a) = \epsilon^{\lambda_1 k} + \epsilon^{-\lambda_2 t_d} \quad (10)$$

ただし，λ_1 (> 0)，λ_2 (> 0) は重みづけ定数である．k は類似記事の密度 m/n，t_d は，

$$t_d = \frac{1}{m}\sum_{i=1,b_i\in\omega}^{m}(t(a) - t(b_i)) \quad (11)$$

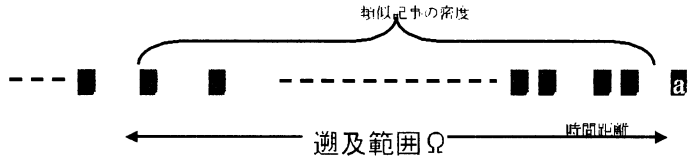

図 3 流行度
Fig. 3 Popularity.

とする．

3.2.3 緊 急 度

放送型ニュース配信システムでは，データはチャンネルという概念に基づいて配信される．それぞれのチャンネルは，あらかじめサーバ側で定義された更新期間（デフォルト更新期間）に従って情報配信を行う．しかし，緊急事態や重要な出来事が発生した場合，チャンネルの更新期間が変更されることがある．たとえば，**図 4** の天気予報チャンネルは，通常は 6 時間ごとに更新されるが，台風が上陸した場合などには，1 時間ごとに更新される．この場合，天気予報チャンネルの配信記事が現時点で価値の高い情報であると考えられる．このような更新頻度（更新期間）の変化をモニタリングすることによって，そのチャンネルの記事の内容の緊急度を評価することができる．

デフォルト更新頻度と比べて，記事 a の所属チャンネル c の新しい更新頻度が高いほどそのチャンネルの記事の緊急度が高くなると考え，記事 a の緊急度 $freq(a)$ を次式のように定義する．

$$freq(a) = \epsilon^{\lambda_3 \sigma_c} \tag{12}$$

$$\sigma_c = \frac{D_c - d_c}{d_c} \tag{13}$$

ただし，σ_c はチャンネル c の更新頻度，D_c は c のデフォルト更新期間，d_c は c の最新の更新期間である．また，λ_3 (> 0) は定数である．

3.3 ユーザプロファイルと時系列的特徴量に基づくフィルタリング

ニュース記事のフィルタリングはユーザプロファイルとニュース記事の時系列的特徴量に基づいて行う．そのフィルタは 3 つの機能を持つ：(1) ユーザプロファイルと記事の類似度計算，(2) チャンネルの更新頻度のモニタリング，(3) 新鮮度・流行度の計算．

まず，配信記事とユーザプロファイルとの類似度を計算する． 基本的には，ユーザプロファイルと高い類似性を持っている記事が選択される可能性が高い．記事 a とユーザプロファイル q の類似度は，記事のキーワードベクトル $v(a)$ とユーザプロファイルのキーワードベクトル $v(q)$ の内積で与える：

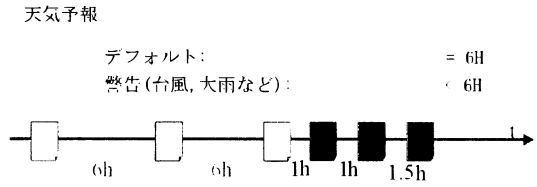

図 4 更新頻度の例
Fig. 4 Example of update frequency.

$$sim(a,q) = \frac{v(a) * v(q)}{|v(a)||v(q)|} \tag{14}$$

ただし，記事のキーワードベクトル $v(a)$ は，ユーザプロファイルのキーワードベクトル $v(q)$ に基づいて生成される．$v(q)$ はユーザがあらかじめ定義したユーザプロファイルのキーワードベクトルとする．

同時に，記事が属するチャンネルの更新頻度をモニタリングする．ほとんどの放送型情報配信システムでは，チャンネルの内容は定期的に更新されるが，突発事件や緊急情報があった場合，更新頻度が変更されることがある．高い更新頻度を持つチャンネルの記事が緊急情報であり，選択されるべきであると考えられる．

さらに，以前の配信情報との類似度/相違点を計算して，記事の新鮮度・流行度を評価する．つまり，記事の内容が新しいか，あるいは以前の記事の続報であるかを評価する．

すなわち，もしある記事がユーザプロファイルとの類似度が高く，高い更新頻度と高い新鮮度・流行度を持っていれば，その記事を選択する．ユーザプロファイルを q とした場合，チャンネル c の記事 a のスコアは次のように計算される．このスコアが閾値より大きい場合，記事 a が選択される．

$$score(a) = \alpha' * sim(a,q) + \beta' * freq(a) + \gamma' * (max(\mu * pop(a), \nu * fresh(a)) \tag{15}$$

ただし，$sim(a,q)$ は，記事 a とユーザプロファイル q との類似度であり，$freq(a)$ は a の緊急度である．$pop(a)$ は，記事 a の流行度である．$fresh(a)$ は a の新鮮度であり，ユーザが $fresh_\Omega(a)$，$fresh_{num}(a)$，$fresh_{cd}(a,\omega)$，$fresh_{dc}(a)$，$fresh_{td}(a,\omega)$ のいずれかを選択することが可能である．また，α'，β'，γ'，μ，

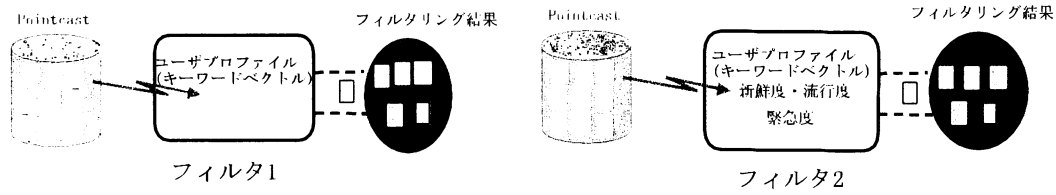

図5 効果の評価実験
Fig. 5 Environment of estimating experiment.

ν は重み付けパラメータであり，ユーザ調整可能である．つまり，ユーザはこれらのパラメータを調整して，自分が重視する情報を得ることができる．たとえば，情報の新鮮度を求めるユーザは新鮮度のパラメータを大きく設定するのに対して，類似度が高い情報を求めるユーザは類似度のパラメータを大きく設定することが可能である．

このフィルタ関数を利用して，ユーザは自分がユーザプロファイルで表現可能な興味に合致した情報，ユーザプロファイルのみでは予測困難な新鮮度の高い情報，および流行度の高い情報を獲得可能である．流行度の高い記事は，最近の話題に関する情報であるので，ユーザにとって価値の高い情報であると思われるが，ユーザプロファイルのみでは獲得は困難である．流行度の高い情報を選択すると同時に，情報の重複を防ぐためにその情報の新鮮度も考える必要があると考えられるので，ここでは，新鮮度と流行度の最大値を記事の選択基準の1つとしている☆．

4. 実験と考察

記事の新鮮度を計算する際，過去の配信履歴をどこまで遡って新鮮度を計算するかによって，結果が異なってくる．本章では，ユーザプロファイルによるフィルタリングとさらにニュース記事の時系列的特徴量を加えたフィルタリングの結果の相異や，新鮮度と遡及範囲の関係について述べる．

4.1 ユーザプロファイルと時系列的特徴量を併用したフィルタリングの効果評価

ユーザプロファイルと時系列的特徴量を併用したフィルタリングの効果を評価するために，Pointcast の1週間のニュース記事を用いて実験を行った（**図5**）．図5では，フィルタ1は従来のユーザプロファイルベースのフィルタリング，フィルタ2は記事の時系列的特徴量（新鮮度・流行度，緊急度）とユーザプロファイルを併用したフィルタリングを行うものである．

実験では，CNN（421件），ZDNet（135件）とSports（512件）の3つのチャンネルの記事を対象としている．特徴量の遡及範囲の記事総数は150件と設定している．フィルタ1のユーザプロファイルと記事の類似度の閾値は，予備実験より0.64に設定した．フィルタ2は3.3節で述べているフィルタ関数（式(15)）を利用している．新鮮度 $fresh(a)$ は統合新鮮度 $fresh_{\Omega}(a)$ を利用し，α，β，γ，δ はそれぞれ0.7，0.1，0.1，0.1と設定している．重み付けパラメータ α'，β'，γ'，μ，ν は，それぞれ5，1，4，1，1に設定している．閾値を4.0に設定している．キーワードはsports，basketball，football，baseball，volleyball，NBA，NHL，MLB，soccor，game，match，Lakers，world，national，team，IT，hardware，software，application，unix，linux，windows，OS，Y2K，virus，database，internet，XML，java，mobileを用いている．

実験の結果を**表1**に示している．ただし，類似度と統合新鮮度および流行度の範囲は $[0,1]$ で，緊急度の基準値は1である．

表1では，全記事とフィルタ2の選択記事の類似度，緊急度と流行度の最小値が同じ値となっている．つまり，フィルタ2はこれら特徴量の最小値をとる記事も選択されている．その原因は，フィルタ2では，1つの特徴量だけでなくすべての特徴量に基づいて情報価値を計算しているためと考えられる．類似度が最小値0であるにもかかわらず選択された記事は，計58件ある．これらの記事の緊急度の平均は2.45，統合新鮮度の平均は0.85，流行度の平均は0.434である．なお，58件のうち，約半分の記事（28件）は，初めに処理を行うときの記事である．これらの記事に対しては，過去の配信記事はない，つまり，遡及範囲の記事数は0であるので，記事の統合新鮮度が高く，選択価値が高く計算されている．

図6に示されているように，フィルタ1は，1068件の記事から729件の記事を選択したのに対して，フィルタ2は，710件の記事を選択した．この2つの選

☆ 新鮮度と流行度は相反する概念であるが，すでに述べたように，新鮮度や流行度の高い情報は，どちらも固定的なユーザプロファイルからは獲得困難である．今回のフィルタリング関数では，流行度と新鮮度の最大値を1つの選択基準にしたが，新鮮度のみまたは流行度のみとしたフィルタリング関数も考えられる．

表 1 効果の評価実験結果
Table 1 Result of estimating experiment.

	全記事集合			フィルタ 1 の結果記事集合			フィルタ 2 の結果記事集合		
	最小値	最大値	平均値	最小値	最大値	平均値	最小値	最大値	平均値
類似度	0	0.810	0.341	0.640	0.810	0.711	0	0.810	0.635
緊急度	0.606	54.5	1.79	-	-	-	0.606	54.5	1.95
統合新鮮度	0.354	1.0	0.673	-	-	-	0.462	1	0.717
流行度	0	0.865	0.235	-	-	-	0	0.865	0.386

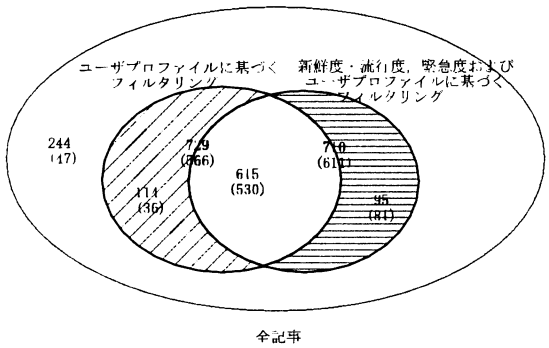

図 6 フィルタリングされた記事集合の関係
Fig. 6 Articles sets: total, filtered by filter 1 and filter 2. The number(x) of articles of each set and the number(y) of valid articles which decided by user are represented as $x(y)$.

択記事集合の交わりは 615 件の記事がある．つまり，キーワードにマッチしている記事 615 件がどちらのフィルタリング手法にも選択されている．Pointcast のジャンルごとに組織されているデータを利用しているので，フィルタ 1 のフィルタ率は高い値となっている．

フィルタ 1 の再現率は 81.6%，適合率は 77.6%である．フィルタ 2 の再現率と適合率はそれぞれ 88.1%，86.1%である☆．時系列的特徴量を考慮して正解記事を選択しているので，再現率と適合率だけではフィルタ 2 はフィルタ 1 より効率を改善しているとはいえないが，フィルタ 2 は新鮮度や緊急度が高い情報を選択できていると考えられる．

フィルタ 1 に選択されるが，フィルタ 2 に選択されない記事は 114 件ある．つまり，キーワードにマッチしている記事の中で，114 件がフィルタ 2 に選択されていない．これらの記事の類似度，統合新鮮度と緊急度の平均は，0.752，0.544，0.915 となっている．その中で，ミス（missed）記事（フィルタ 2 に選択されなかった正解記事）は 36 件であった．

フィルタ 1 に選択されないが，フィルタ 2 に選択される記事は 95 件あり，類似度，統合新鮮度および緊急度の平均が 0.224，0.839，38.251 である．つまり，キーワードにマッチしていない記事の中で 95 件がフィルタ 2 に選択されている．この中の失敗（failured）記事（フィルタ 2 に選択された不正解記事）は 14 件であった．

理想では，式 (16)，(17) で定義しているミス率 r_m（missing ratio）と失敗率 r_f（failure ratio）はともに 0 であるが，今回の実験では，ミス率 r_m と失敗率 r_f はそれぞれ 31.6%，14.7%となる．つまり，ユーザプロファイルのみのフィルタリング手法と比べて，記事の時系列的特徴量とユーザプロファイルを併用したフィルタリング手法は，ミス率 31.6%で情報の重複が減少し，失敗率 14.7%で緊急度または新鮮度・流行度が高い情報を追加している．

ミス率 r_m と失敗率 r_f は次のように計算される．

$$r_m = \frac{s_M}{s_{T-T\cap V}} \tag{16}$$

$$r_f = \frac{s_F}{s_{V-V\cap T}} \tag{17}$$

ただし，M はミス記事の集合，T はユーザプロファイルのみのフィルタリング手法によって選択される記事集合，V は記事の時系列的特徴量とユーザプロファイルを併用したフィルタリング手法の選択記事集合，F は失敗記事集合である．s_X は集合 X のサイズを表す．

4.2 新鮮度と遡及範囲

新鮮度の計算には，様々な方式がある．この中で，遡及範囲の記事数の変化によって新鮮度が変化するのが一般であるが，ユーザが遡及範囲を決めることが困難な場合，遡及範囲に依存しない新鮮度の計算方式が求められる．特に，リアルタイム処理が要求されるオンラインニュース配信システムでは，計算量を抑えることが非常に重要なので，遡及範囲の影響を緩和した新鮮度の計算が必要である．以下に示すように，本論文で提案している統合新鮮度の計算モデルは，適切なパラメータを選択することで，遡及範囲の変化による新鮮度への影響をほとんどなくすことができ，安定した新鮮度を求めることが可能である．Pointcast の

☆ 正解記事を選択するとき，キーワードマッチしているかのほかに，内容が新しいか，最近の話題であるか，緊急性があるかなども選択基準とした．

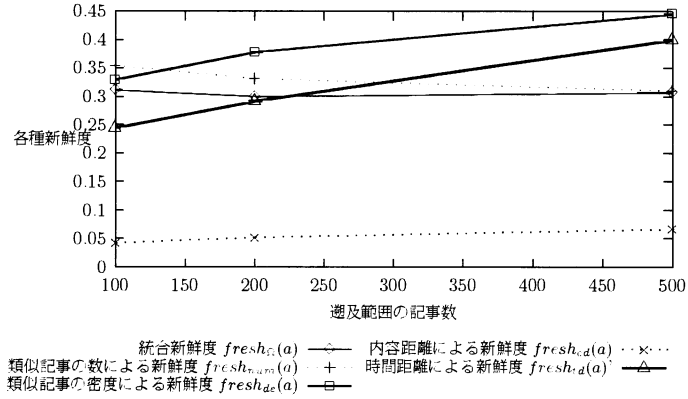

図 7 記事総数で遡及範囲を決める場合
Fig. 7 Freshness and retrospective scope by total number of articles.

表 2 記事の数で遡及範囲を決める場合
Table 2 Freshness and retrospective scope by total number of articles.

No.	全記事数	類似記事数	時間幅	統合新鮮度
I-1	100	7.90	58.71	0.311
I-2	200	11.22	115.13	0.303
I-3	500	19.27	313.11	0.306

CNN チャンネルの約 1 カ月分の配信記事 1014 件を用いて実験を行った結果，その効果を確認している．

実験 I：記事数の総数で遡及範囲を決める場合

実験 I では，遡及範囲の記事総数を 100，200，500 にして 3 回の実験を行った．**表 2** で示しているように，1 回目の実験では，遡及範囲の記事は 100 件，平均類似記事は 7.90 件，遡及範囲の時間幅の平均は 58.71 時間，統合新鮮度の平均は 0.311 である．2 回目の実験では，遡及範囲の記事は 200 件，平均類似記事は 11.22 件，遡及範囲の時間幅の平均は 115.13 時間，統合新鮮度の平均は 0.303 である．3 回目の実験では，遡及範囲の記事は 500 件，平均類似記事は 19.27 件，遡及範囲の時間幅の平均は 313.11 時間，統合新鮮度の平均は 0.306 である．

図 7 では，各新鮮度（類似記事の数，類似記事の密度，時間距離と内容距離による新鮮度）および統合新鮮度と遡及範囲の関係を示している．遡及範囲と，統合新鮮度との相関係数は -0.559，類似記事の数による新鮮度との相関係数は -0.996，類似記事の密度による新鮮度との相関係数は 0.999，内容距離による新鮮度との相関係数は 0.998，時間距離による新鮮度との相関係数は 0.991 である．遡及範囲と新鮮度の相関が高いことが明らかである．

遡及範囲の増大にともなって，新鮮度を増大させるのは類似記事の密度，内容距離および時間距離で，減少させるのは類似記事の数であることが分かる．遡及範囲の記事総数の変化にともなって，新鮮度に与える影響が一番大きいのは類似記事の密度である．統合新鮮度は，遡及範囲による影響を緩和していることが分かる．

実験 II：類似記事の数で遡及範囲を決める場合

実験 II では，類似記事の数を 10，20 にして 2 回の実験を行った．**表 3** で示しているように，1 回目の実験では，遡及範囲の記事は平均 92.63 件，平均類似記事は 9.55 件，遡及範囲の時間幅の平均は 49.41 時間，統合新鮮度の平均は 0.229 である．2 回目の実験では，遡及範囲の記事は平均 155.28 件，平均類似記事は 18.26 件，遡及範囲の時間幅の平均は 78.81 時間，統合新鮮度の平均は 0.205 である．実際の類似記事の数が設定値と違うのは，処理記事の中で，以前配信した記事（実験の記事範囲内）での類似記事が設定値より少ないものがあるからである．

図 8 から，時間距離，類似記事の密度と内容距離は，遡及範囲の増加とともに新鮮度を増加させるのに対して，類似記事の数は新鮮度を減少させる働きがあることが分かる．遡及範囲は，時間距離による新鮮度への影響が一番大きい．統合新鮮度は，遡及範囲の増大にともなって，緩やかに減少していることが分かる．

実験 III：時間幅で遡及範囲を決める場合

実験 III では，遡及範囲が時間幅で決められる．時間幅を 72 時間，144 時間と 240 時間にして，3 回の実験を行った．**表 4** で示しているように，1 回目の実験では，遡及範囲の記事は平均 169.32 件，平均類似記事は 10.68 件，遡及範囲の時間幅の平均は 80.13 時間，統合新鮮度の平均は 0.304 である．2 回目の実験では，遡及範囲の記事は平均 221.21 件，平均類似記事は 12.84 件，遡及範囲の時間幅の平均は 170.05 時

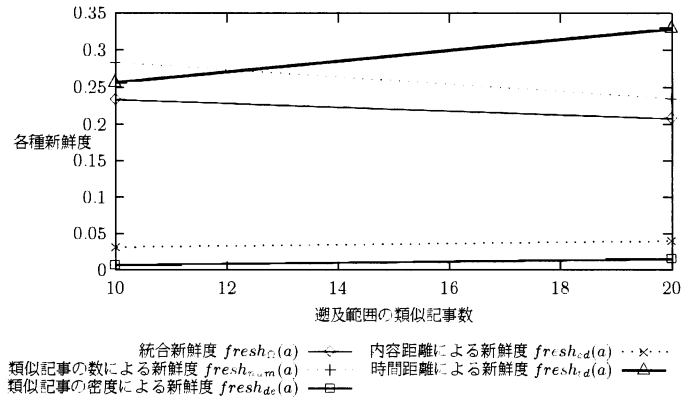

図 8 類似記事数で遡及範囲を決める場合
Fig. 8 Freshness and retrospective scope by number of similary articles.

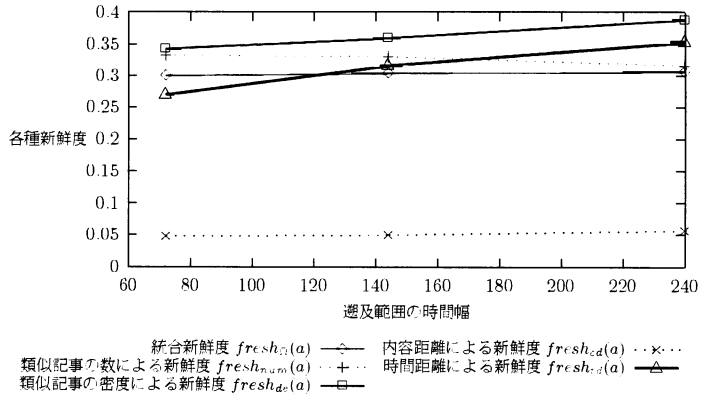

図 9 時間幅で遡及範囲を決める場合
Fig. 9 Freshness and retrospective scope by time interval.

表 3 類似記事の数で遡及範囲を決める場合
Table 3 Freshness and retrospective scope by number of similary articles.

No.	全記事数	類似記事数	時間幅	統合新鮮度
II-1	92.63	9.55	49.41	0.229
II-2	155.28	18.26	78.81	0.205

表 4 時間幅で遡及範囲を決める場合
Table 4 Freshness and retrospective scope by time interval.

No.	全記事数	類似記事数	時間幅	統合新鮮度
III-1	169.32	10.68	80.13	0.304
III-2	221.21	12.84	170.05	0.306
III-3	272.60	13.81	252.01	0.307

間，統合新鮮度の平均は 0.306 である．3 回目の実験では，遡及範囲の記事は平均 272.60 件，平均類似記事は 13.81 件，遡及範囲の時間幅の平均は 252.01 時間，統合新鮮度の平均は 0.307 である．実際の時間幅が設定値と一致していないのは，放送型情報配信システムでは，配信時間は固定でないからである．

図 9 では，遡及範囲と，統合新鮮度との相関係数は 0.980，類似記事の数による新鮮度との相関係数は −0.947，類似記事の密度による新鮮度との相関係数は 0.997，内容距離による新鮮度との相関係数は 0.984，時間距離による新鮮度との相関係数は 0.988 である．遡及範囲の変化は，類似記事の密度による計算される新鮮度への影響が大きいことが分かる．また，統合新鮮度は，遡及範囲の影響を緩和していることも分かる．

流 行 度

新鮮度と同じく，流行度も遡及範囲に依存する．**図 10**，**図 11**，**図 12** では，遡及範囲をそれぞれ記事総数，類似記事数，時間幅で決める場合の流行度の実験結果を示している．図から，流行度も統合新鮮度と同じく，安定した値をとることができることが分かる．

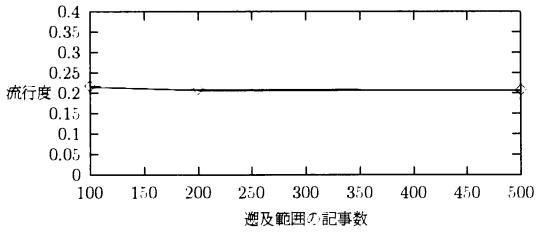

図 10 流行度と遡及範囲：記事総数の場合
Fig. 10 Popularity and retrospective scope by total number of articles.

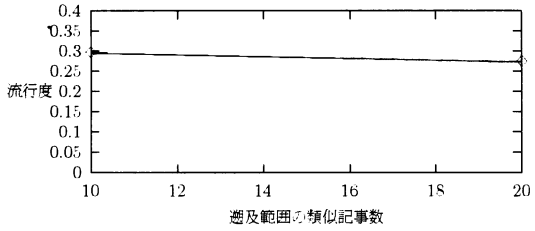

図 11 流行度と遡及範囲：類似記事数の場合
Fig. 11 Popularity and retrospective scope by number of similar articles.

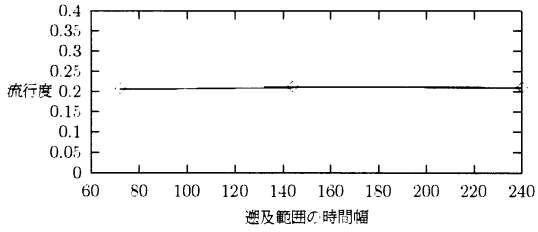

図 12 流行度と遡及範囲：時間幅の場合
Fig. 12 Popularity and retrospective scope by time interval

考　　察

実験 I から，遡及範囲は記事総数で決める場合では，記事総数がある一定の値まで増加するに従って，統合新鮮度が上がるが，この値を超えると，統合新鮮度が下がる傾向があることが分かる．

実験 II から，時間間隔を用いて遡及範囲を決める場合，遡及範囲の拡大にともなって，統合新鮮度が増大する傾向があるが，時間間隔が大きくなると，統合新鮮度の変動幅は小さくなることが分かる．

実験 III から，類似記事の数を用いて遡及範囲を決める場合では，類似記事の数，つまり遡及範囲の増加に従って，統合新鮮度が減少していることが分かる．

類似記事の密度によって計算された新鮮度は，実験 II の結果が，実験 I と実験 III の値に比べて，はるかに小さくなっている．同時に，新鮮度の変動幅は，他の実験（I と III）より大きくなっている．したがって，類似記事の密度は，新鮮度の変動幅への影響が大きいと考えられる．これらの実験結果から，以下のことが確認できた．

- 遡及範囲は，類似記事の密度および時間距離により計算される新鮮度に及ぼす影響が大きい．
- 遡及範囲の増大に従って新鮮度を増大させるのは類似記事の密度，時間距離および内容距離である．一方，減少させるのは類似記事の数である．
- 類似記事の密度は，新鮮度の変動幅に与える影響が大きい．

遡及範囲の変化にともなって，計算される新鮮度が異なるが，類似記事の数，類似記事の密度，時間距離と内容距離を綜合して考慮した統合新鮮度は，遡及範囲の変化の影響を緩和することが可能である．しかし，今回の実験データの量が少ないため，十分な確認ができたとはいえないので，今後はデータの量を増やしてさらに実験を行う必要があると思われる．

5. おわりに

ユーザが自分の欲しい情報を探しにいくのではなく，あらかじめサーバ側で登録されているユーザプロファイルを用いて，新しい情報をユーザに自動的に配信する放送型情報配信システムが注目されている．

大量の配信情報から，ユーザが興味ある情報を受信できるようにするためには，情報フィルタリングや情報検索といった情報処理が必要となる．しかし，従来の情報フィルタリング手法では，配信情報の時系列性を配慮していない．よって，重複した情報配信や情報の漏れなどの問題が起こる場合がある．

本論文では，配信記事の時系列性を考慮して，放送型情報配信システムにおけるニュース記事の特徴量を，配信情報の更新頻度，配信内容，配信履歴などに基づいて定義し，これに基づいて，ユーザプロファイルと併用した情報フィルタリング手法を提案した．

本論文で提案した情報フィルタリング手法の効果をまとめると，以下のとおりである：

- 情報の重複配信を防ぐとともに，より新しく，重要な情報をユーザに配信することが可能である．
- 類似記事の数，類似記事の密度，時間距離と内容距離を総合して考慮した統合新鮮度は，遡及範囲の増減による影響を緩和することが可能である．

記事の時系列的特徴量とユーザプロファイルを併用した情報フィルタリング手法は，従来のキーワードのみの情報フィルタリング手法と比べて，ユーザはより新しく，重要な情報を獲得可能であることが実験結果によって証明された．すなわち，ユーザは自分がユーザ

プロファイルで表現可能な興味に合致した情報，ユーザプロファイルのみでは予測困難な新鮮度の高い情報，および流行度の高い情報を獲得可能である．

記事の新鮮度は，遡及範囲，つまり配信履歴をどこまで遡って計算するかによって異なる結果になるが，本論文で提案されている統合新鮮度の計算モデルは，遡及範囲の新鮮度に与える影響を緩和することが可能であることが実験結果によって確認できた．

今後は，実験データの量を増やしてさらに実験を行い，ニュース記事の時系列的特徴量の抽出方法とフィルタリング関数の改善を行う予定である．同一チャンネルで異なる遡及範囲での情報フィルタリングの安定性と，複数のチャンネルでの振舞いの違いなどを検証することも予定している．また，適合フィードバックなどによって，時系列的特徴量のパラメータの自動決定手法などの課題を取り上げたいと考えている．

謝辞 本研究の一部は，文部省科学研究費「分散型ハイパーメディアからの構造発見とアクセス管理」（課題番号は「12680416」）の援助を受けています．また，本研究の一部は，日本学術振興会未来開拓学術研究推進事業における研究プロジェクト「マルチメディア・コンテンツの高次処理の研究」（プロジェクト番号 JSPS-RFTF97P00501）によっています．ここに記して謝意を表します．

参 考 文 献

1) 角谷和俊，宮部義幸：放送型情報配信のためのモデルとシステム，情報処理学会論文誌：データベース，Vol.40 No.SIG 8(TOD4), pp.141–157 (1999).
2) Acharya, S., Alonso, R., Franklin, M. and Zdonik, S.: Broadcast Disks: Data Management for Asymmetric Communication Environment, *Proc. ACM SIGMOD '95*, pp.199–210 (1995).
3) Acharya, S., Franklin, M. and Zdonik, S.: Balancing Push and Pull for Data Broadcast, *Proc. ACM SIGMOD '97*, pp.183–194 (1997).
4) Aksoy, D., Altinel, M., Bose, R., Cetintemel, U., Franklin, M. and Zdonik, S.: Research in Data Broadcast and Dissemination, *Proc. 1st International Conference on Advanced Multimedia Content Processing (AMCP'98)*, pp.196–210 (1998).
5) Franklin, M. and Zdonik, S.: "Data In Your Face": Push Technology in Perspective, *Proc. ACM SIGMOD '98*, pp.516–519 (1998).
6) Gerwig, K.: The Push Technology Rage... So What's Next?, *ACM netWorker: The Craft of Network Computing*, Vol.1, No.2, pp.13–17 (1997).
7) 馬 強，角谷和俊，田中克己：放送型情報配信システムにおける仮想チャネルとその XML による実現，電子情報通信学会第 10 回データ工学ワークショップワークショップ論文集（DEWS'99）(1999).
8) Ma, Q., Kondo, H., Sumiya, K. and Tanaka, K.: Virtual TV Channel: Filtering, Merging and Presenting Internet Broadcasting Channels, 情報処理学会研究会報告，Vol.99, No.61, pp.189–194 (1999).
9) Ma, Q., Kondo, H., Sumiya, K. and Tanaka, K.: Virtual TV Channel: Filtering, Merging and Presenting Internet Broadcasting Channels, *Proc. ACM Digital Library Workshop On Organizing Web Space (WOWS)* (1999).
10) Wong, J.: Broadcast Delivery, *Proc. IEEE*, Vol.76, No.12 (1988).
11) PointCast, http://www.pointcast.com (1999).
12) Ramakrishnan, S. and Dayal, V.: The PointCast Network, *Proc. ACM SIGMOD '98*, p.520 (1998).
13) Yan, T.W. and Garcia-Molina, H.: SIFT – A Tool for Wide-Area Information Dissemination, *Proc. 1995 USENIX Technical Conference*, New Orleans, Louisiana, USA, USENIX Association (Ed.), pp.177–186 (1995).
14) Shapiro, D.M.: Push-Based Web Filtering Using PICS Profiles, http://www.w3.org/TandS/Public/Theses/DavidShapiro/thesis-dshapiro.html (1998).
15) 武田皓一，中村祐一，浦本直彦：XML がもたらす創造的ネットワーク－動的な情報源と分散エージェント，人工知能学会誌，Vol.14, No.6, pp.35–43 (1999).
16) Takeda, K. and Normiyama, H.: Site outlineing, *Proc. ACM Digital Libraries'98*, pp.309–310, ACM Press (1998).
17) Kamba, T., Sakagami, H. and Koseki, Y.: Automatic personalization on push news service, W3C Push Workshop, http://www.w3.org/architecture/9709_Workshop/paper02/paper02.html (1997).
18) Allan, J., Papka, R. and Lavrenko, V.: On-line New Event Dectection and Tracking, *Proc. SIGIR'98*, pp.37–45 (1998).
19) Yang, Y., Pierce, T. and Carbonell, J.: A Study on Retrospective and On-Line Event Detection, *Proc. SIGIR'98*, pp.28–36 (1998).
20) 宗像浩一，吉川正俊，植村俊亮：鮮度と同期度に基づく周期データの選択方式，アドバンスト・データベース・シンポジウム'99（ADBS'99），pp.141–150 (1999).

21) Salton, G.: *Automatic Information Organization and Retrieval*, McGraw-Hill (1968).

(平成 12 年 3 月 20 日受付)
(平成 12 年 6 月 17 日採録)

(担当編集委員 諸橋 正幸)

馬 強(学生会員)

2000 年神戸大学大学院自然科学研究科博士前期課程修了. 現在, 同大学大学院自然科学研究科博士後期課程在学中. データベース, 放送型情報メディアに興味を持つ.

角谷 和俊(正会員)

1988 年神戸大学大学院工学研究科修士課程修了. 同年松下電器産業(株)入社. ソフトウェア開発環境, マルチメディアデータベース, データ放送の研究開発に従事. 1998 年神戸大学大学院自然科学研究科博士後期課程(情報メディア科学専攻)修了. 1999 年神戸大学都市安全研究センター都市情報システム研究分野(工学部情報知能工学科兼担)講師. 博士(工学). 情報処理学会データベースシステム研究会幹事. ACM, IEEE Computer Society, システム制御情報学会各会員.

田中 克己(正会員)

1974 年京都大学工学部情報工学科卒業. 1976 年同大学大学院修士課程修了. 1979 年神戸大学教養部助手, 1986 年同大学工学部助教授. 1994 年同大学工学部教授(情報知能工学専攻). 1995 年同大学大学院自然科学研究科(現在, 情報メディア科学専攻)専任教授, 現在に至る. 工学博士. 主にデータベースの研究に従事. 人工知能学会, IEEE Computer Society, ACM 等会員.