WebSCAN：Webサイトの変更発見と放送型変更通知

宮崎慎也[†] 馬強[††] 田中克己[†††]

Web上には大量の情報が溢れ，つねにその量は増加している．ユーザがつねにその変更の中から価値ある変更を探し出すことは非常に困難な作業である．本論文では，Webサイトを監視し，サイト内のWebページの変更，追加に対して，その価値を内容，サイトの構造などに基づいて解析し，ユーザへの変更通知を行う機構について述べる．変更ページの解析手法として，その価値を評価するための尺度として新鮮度，流行度，アクセス頻度，更新頻度などを用いる．また，変更通知機構として，変更データをプッシュ型配信機構により配信する．変更データには解析結果の情報が含まれ，受信された変更データのすべてを呈示するのではなく，ユーザごとのユーザ・プロファイルによりパーソナライズされ，各ユーザにとって価値のある変更情報に変換，呈示する．

WebSCAN: Content-based Change Discovery and Broadcast-notification for Web Sites

Shinya Miyazaki,[†] Ma Qiang[††] and Katsumi Tanaka[†††]

The vast amount of information is availabel on the WWW, and grows rapidly. It's not easy for user to acquire the valuable new information via the Internet. In this paper, we propose a change monitoring/notification system *WebSCAN* (Web Sites Change Analyzer and Notifier) for Web sites, which monitors and analyzes the changes of Web sites to notify a user the important changes by a push-type delivery mechanism. In WebSCAN, the changes of Web sites are not only monitored periodically, but also are estimated by the content, browsing frequency and update frequency. The structure of Web sites is also considered to estimate the change worth at WebSCAN. Based on the estimated *change worth*, the notification is generated and delivered to user automatically with the push technology.

1. はじめに

近年，WWW（World Wide Web）上の情報は大量かつ複雑なものとなっている．利用者はブラウジングや検索エンジンの利用，またはWebブラウザのブックマークに登録されている好みのサイトにアクセスし，そこから新たな情報を得ようとする．しかしながら，Webの世界はつねに変化し，毎日のように大量の新しいWebページが作られる．逆にいつ，どのようにWebページが更新・追加されるかもユーザには分からない．ユーザにとって，興味あるWebサイト，ページを欠かさず閲覧し続けることは，非常に困難な場合がある．一方で，Webのよう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[2]が提案されている．これはユーザのプロファイルに従って，ユーザが受動的に情報を受信できるため，ユーザの情報を獲得するための負担が軽減される利点がある．

また現在，Webの変更通知に関してWebCQ[15),16)]，NetMind[11)]など，いくつかの既存システム，アプリケーションがあるが，これらの多くは単にWeb上の変更の監視，通知を行うのみであって，どの変更が重要なのかという変更に対する解析を行うものではない．

そこで我々は，本論文においてWebサイトに対する変更監視・解析・通知システム—WebSCAN（Web Sites Change Analyzer and Notifier）—を提案する．これは，Webサイトを定期的に監視し，発見した変更の価値を内容，更新時間，構造およびページ（サイト）へのユーザアクセス頻度などの解析をもとに推測する．ここで考える変更とはWebサイト内のWebページ追加・更新であるが，推測されたページの価値をもとに，放送型情報配信に基づいてユーザに有用と思われ

† 神戸大学大学院自然科学研究科情報知能工学専攻
Division of Computer and System Engineering, Graduate School of Science and Technology, Kobe University
†† 神戸大学大学院自然科学研究科情報メディア科学専攻
Division of Information and Media Science, Graduate School of Science and Technology, Kobe University
††† 京都大学大学院情報学研究科社会情報学専攻
Division of Social Informatics, Graduate School of Informatics, Kyoto University

る変更のみを通知するシステムである．

既存システムと比べ，WebSCAN の特徴を以下に述べる．

- **コンテンツと構造を考慮した変更解析**

  複数の変更の中から価値ある変更を見つけ出すために，WebSCAN は変更コンテンツと Web サイトの構造の両者に基づいて変更解析を行う．変更の価値を評価するために，WebSCAN では，Web ページ（サイト）の構造に基づいて比較対象を選択して，変更前後のコンテンツの比較を行う．

- **変更の内容に対する，意味的な解析**

  変更，追加ページが，それ以前のページやその周辺に存在するページとの類似性が低い場合，その情報は，既存の情報とは似ていない，今までにはあまり存在しない新鮮度（freshness）の高い情報である可能性が考えられる．また，その類似性が高い場合，その情報は既存の情報と似通った情報であり，そのサイト内ではよく存在する，流行度（popularity）の高い情報である可能性が考えられる．

  情報の新鮮度や流行度は，人間の経験などに基づくものであり，それを客観的に評価することが困難である．それゆえに挑戦しがいのあるテーマであると考える．本論文では，ページ（サイト）の変更前後の内容，更新時間およびページ（サイト）の構造などに基づいて新鮮度・流行度を計算する試みを行う．提案する新鮮度・流行度の計算手法では，内容の類似・非類似性が重要な尺度となっている．内容の類似性については情報検索の分野で長い研究の歴史があり，WebSCAN では，コサイン相関値という伝統的な手法を採用して内容の類似・非類似性を計算している．本論文の目的は，類似検索における「類似度」をどのように定義すべきかの研究ではなく，伝統的な文献の類似度をもとにして，文書間の新鮮度，流行度を測る尺度を提案し，これに基づいて，Web サイトにおける「重要」な変更を利用者に通知する方式を開発することである．

- **プッシュ型変更通知とパーソナライゼーション**

  変更に対する価値評価などを含めた変更データを作成し，クライアントへ配信する．各クライアント側では，ユーザの興味に基づくユーザ・プロファイルにより，配信された変更情報から各ユーザ独自の変更情報を動的に作成し，呈示する．

表 1 に我々が提案するシステム WebSCAN と，WebCQ [15] などに代表される既存システムとの比較を示す．既存のシステムやアプリケーションの特徴として，ユーザの指定した単独ページを監視することや，またページ内の監視する対象（フレーズ，画像，リンクなど）を細かく指定できるなどがあげられる．そのためこれらのシステムは，たとえばこのページに画像が追加されたら知りたい，この商品の値段が変更されたら知りたい，このページのリンク先が増えたら知りたいなど，ユーザが何らかの明確な監視目的を持っている場合に有用ではないかと考えられる．これに対し，我々の提案するシステムでは，ユーザが指定した Web サイトが監視対象であり，ユーザに対し，そこに新しく追加されたページや，更新されたページの解析とともに内容を通知・呈示することを目的とする．

表 1 システム比較
Table 1 WebSCAN vs. conventional system.

	既存システム	WebSCAN
監視対象の粒度	小 Web ページ	大 Web サイト
監視対象の指定	可	可
変更箇所と他の部分との比較	無	有
変更の意味的な解析	無	有

以下，2 章で本研究の基本コンセプトについて述べ，3 章で変更価値を決定するための評価手段と，実験によるその評価を行う．4 章ではプッシュ型配信とユーザ・プロファイルによる変更通知のパーソナライゼーションについて述べる．5 章ではプロトタイプシステム WebSCAN について，6 章では本研究の関連研究について触れ，7 章に本論文のまとめを述べる．

2. 基本コンセプト

WebSCAN の基本コンセプトを**図 1** に示す．サーバ側で Web サイトを監視し，そこで発生した変更に対して新鮮度/流行度，アクセス頻度，更新頻度といった尺度で解析を行う．それらの情報をもとに各変更ページに関する情報をプッシュ型配信によって各ユーザ（クライアント）に配信する．クライアント側では受信した変更情報から各ユーザのユーザ・プロファイルを用いることで，独自の変更情報が呈示される．

変更の価値を計算するために，WebSCAN は，まず変更後のコンテンツと変更前のコンテンツの類似・非類似値を計算する．ここで，類似度の高い場合，前の情報の補充，続報である可能性が高いと考えられる．ユーザの興味ある話題であれば，このような続報情報は変更価値が高いと考えることができる．一方，類似していない，すなわちその非類似性が高い場合ほど，ユーザにとっては見慣れない，新しい情報であると考

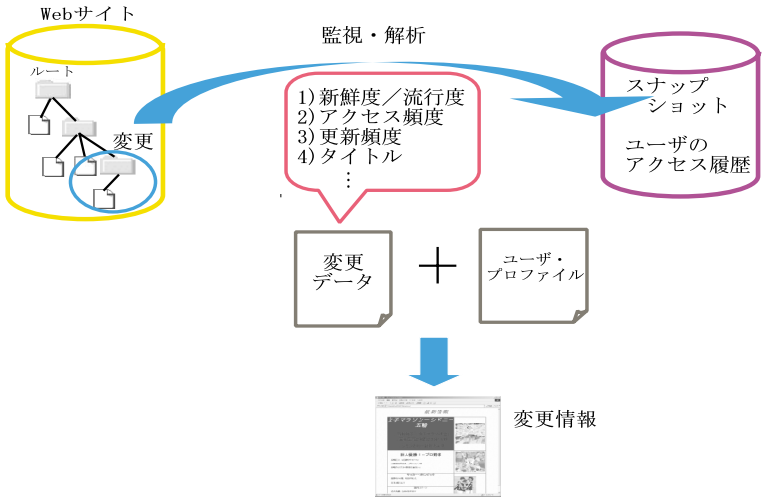

図 1 基本コンセプト
Fig. 1 Basic concept of WebSCAN.

えられ，その価値は高いと考えることができる．そこで内容の類似/非類似性を考慮する際に，類似度，非類似度をもとに計算される新鮮度/流行度という尺度を用いる．

ユーザのアクセス頻度は，ユーザの興味を反映していると考えられる．すなわちアクセス頻度が高いページ，トピックの内容ほど，ユーザが興味のある，気に入っている内容であると考えられる．このようにユーザが興味のあるページ，トピックほど，変更価値があると考えられる．

ページの更新頻度もまた，変更の価値に影響すると考える．すなわち毎日定期的に更新されるページと，数カ月ぶりに更新されたページとでは，一般に後者の変更の方が，ユーザに通知する価値があると考えられる．つまり更新頻度の低いページの変更ほど，変更の価値が高いと考えられる．

これら尺度を用いて，変更に対する価値を評価する．そして，これらの解析結果を含めた，各変更ページに関する情報が自動作成され，ユーザに配信される．

また，各ユーザごとに興味，関心のあるトピックは大きく違い，また変更の価値も大きく違う．ユーザに通知する際に，各ユーザごとにパーソナライズされた変更通知が望まれる．そこで，ユーザの興味を重視して，重要な変更を抽出するために，ユーザ・プロファイルを利用する．WebSCAN では，変更に関する情報を各クライアントにプッシュ型配信し，クライアント側において，配信された情報から，各ユーザ独自の変更情報をフィルタリングし，変更通知の呈示を行う．

3. Web に対する変更解析と評価

Web 上の情報が変更された場合☆に，その情報の価値を，内容の類似・非類似性，アクセス頻度，更新頻度などに基づいて推測する．

3.1 比較範囲

変更内容の類似・非類似性などを評価するためには，変更の種類に応じて適切な比較範囲が必要である．ここでは，まず変更を評価する際の対象となる比較範囲について述べる．

変更の種類として replace と new の 2 通りを考える．

- replace は，ページ内の部分的な変更，またはページ単位の置き換えである．この場合，変更前後でその内容に対する比較を行う．
- new は，ページに対する新たな内容の追加，または新たなページの生成である．この場合，追加された内容に対して，他の部分，他のページとの比較を行う．

変更ページに対して，Web サイトあるいはページの構造に基づいて，変更部分と同じレベルのその他の部分をその比較範囲として選択する．それを以下に説明する．

3.1.1 ページ内の変更

既存ページ内に変更，あるいは追加が生じた場合，**図 2** のように比較範囲を選択する．変更の場合は，変更前の内容との比較，また追加の場合は追加前の内容，

☆ 本論文では Web ページ（HTML 文書）を対象に，その変更について考える．

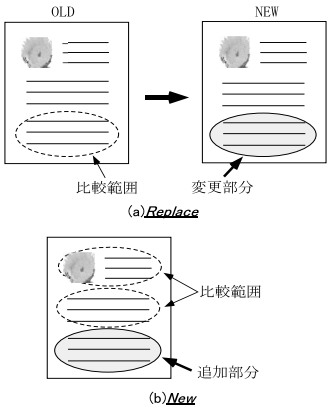

図 2 比較範囲：ページ内の変更
Fig. 2 Comparison scope: case of page modification.

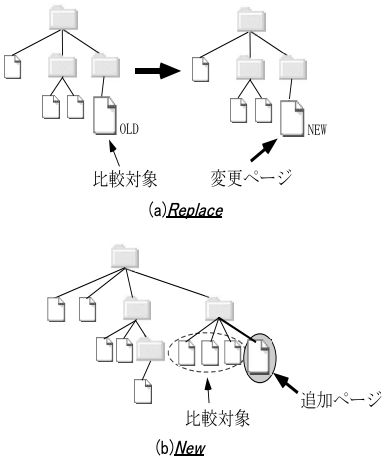

図 3 比較範囲：ページ単位の変更
Fig. 3 Comparison scope: case of new page.

つまりページのその他の部分との比較を行う．これは追加された部分を 1 つの固まり（パラグラフ）ととらえ，HTML の DOM（Document Object Model）に基づく木構造から，その他の部分を同じ階層レベルに切り分け，それぞれを比較対象とする．つまり文書の構造に基づいて比較範囲を選択する．

3.1.2 ページ単位の変更

新しいページの生成は Web 上で頻繁に行われる変更の 1 つであると思われる．ページ単位で変更（書き換え），あるいは追加が生じた場合，**図 3** のように比較範囲を選択する．replace の場合は，変更前のページとの比較を行う．new すなわち追加の場合，追加されたページと同一ディレクトリ以下に格納されているページを比較範囲とする．また，サブディレクトリ以下に含まれる各ページも比較対象とする．

変更ページの新鮮度・流行度は，類似/非類似計算

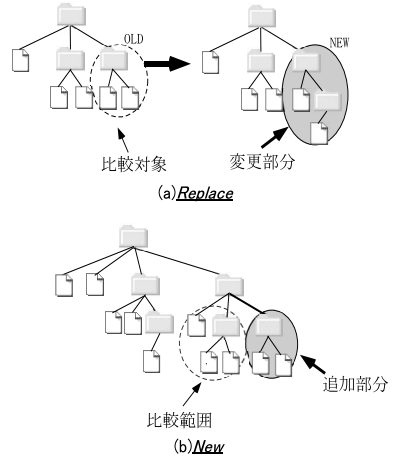

図 4 比較範囲：複数ページの変更
Fig. 4 Comparison scope: case of new topic.

に基づくが，同じトピックに属するページと比較することがより公平であると考えられる．Web サイトの 1 つのディレクトリ内のページ群は，何らかの共通の話題，トピックを形成している可能性が高いと考えられるので，関連する Web ページは同じディレクトリ（URL パス）内に格納されている可能性が高い．したがって，Web サイトのディレクトリ構造（木構造）から，同じトピックに属するページを選択して変更ページの比較範囲とすることが考えられる．Web の構造に関する研究[3),4)] の多くは，リンク解析によって得られたグラフ構造に基づいて行っている．我々は，リンクではなく，Web ページの URL パスの解析に基づいて Web サイトの構造を木構造，つまりディレクトリ構造として抽出している．

3.1.3 複数ページの変更

複数のページの集まりがトピックとして，変更，あるいは追加された場合，**図 4** のように比較範囲を考える．変更トピックを単位として，3.1.2 項と同様に比較範囲を選択する．replace の場合は，変更前のトピックを比較対象とし，また new の場合は，トピック（ディレクトリ）を単位として，同じディレクトリに含まれるページ，トピックを比較範囲とする．

3.1.4 マルチサイト間での解析

異なる Web サイトが同じような情報を提供している場合がある．たとえば，ある大事件が発生すると，いくつものニュースサイトで同じ情報が提供される．このように関連ある複数の Web サイトを考慮して，変更の価値を評価することも考えられる．このような場合，**図 5** に示すように関連する複数のサイト間で，あらかじめ類似するトピック（ページの集合）を発見し

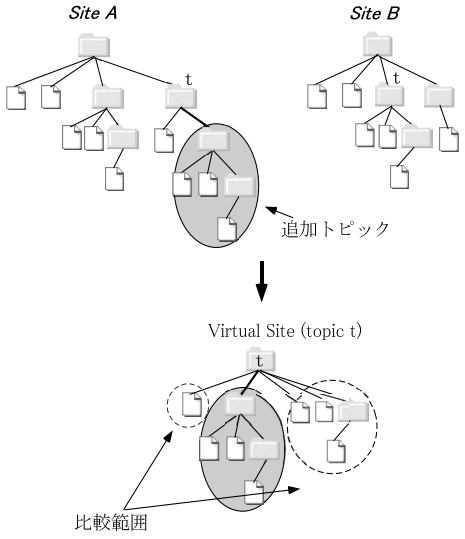

図 5 比較範囲：複数サイト間での比較
Fig. 5 Comparison scope: case of related Web sites.

ておく．あるサイトで変更が生じた場合，変更が含まれるトピックに対して，それに類似する他サイトのトピックを含めて仮想サイトを構築する．仮想サイト内の変更，あるいは追加トピックに対して，同一のディレクトリ内に含まれるページ，トピックを比較範囲として設定する．

3.2 変更の価値

変更に対して時間，構造およびコンテンツに基づいて変更の価値を評価する．変更の価値を計算する要素として内容の類似・非類似性，アクセス頻度，更新頻度をあげる．たとえば新しい情報が，他のページの内容あるいは変更前の内容に対し，その類似性が低い場合は，その内容は新規性の高いという意味で価値ある変更と考えることができる．また類似性が高く，それらの時間的な距離が小さい場合，その内容は流行の度合いが高いという意味で価値ある変更と考えることができる．ここで我々は新鮮度，流行度という尺度を用いて，変更内容の価値を推測する．また，アクセス頻度の高いページ，トピックでの変更は，一般的にその価値が高いと考え，更新頻度の低いページの変更ほど，ユーザに通知する価値が高いと考える．

3.2.1 新鮮度/流行度

情報の新しさに関する研究は近年注目されるようになってきている．その代表としては，Georigia 大学の Infosphere プロジェクト[5]，Stanford 大学の WebBase プロジェクト[6),7)]，奈良先端大の宗像ら[8)] などがある．これらの研究では，主に時間に基づいて情報の新しさ（freshness）を求めているが，馬ら[1)] は時間とコンテンツの両方を考慮して時系列データに対する情報フィルタリングのための尺度（新鮮度/流行度）を提案している．馬らは，過去の履歴に基づいて新鮮度，流行度を計算しているが，我々は馬らの定義をベースに，Web サイト（ページ）の構造に基づいて類似・非類似の比較対象を選択して新鮮度と流行度の計算を行う．

Web サイトは，時間の経過とともに更新される．したがって，Web ページは時系列文書として見なすことができる．3.1 節では，ページの変更価値を計算するための比較範囲が Web サイトの木構造に基づいて選択されている．この比較範囲を時間軸から見ると，変更前のページ（Web サイト）群である．つまり，変更前後のページ（Web サイト）と比較し，類似・非類似の計算に基づいて新鮮度・流行度といった変更価値を計算している．

新 鮮 度

変更されたページ，あるいは追加されたページの内容が，変更前もしくは他のページとまったく違う内容であるとき，それは価値ある変更であるといえる．ここで我々は新鮮度という尺度を次のように定義する．

あるページ A が変更，追加された場合，ページ A に対する新鮮度は以下のように与えられる．

(1) 類似ページ数による新鮮度

ページ A に対する類似ページ数が少ないほど，ページ A は価値が高い．そこで類似ページ数による新鮮度を次のように与える．

$$fresh_{num}(A) = \frac{1}{log_2(2+m)} \tag{1}$$

ここで，m は比較範囲☆ Ω 内における，A の類似ページ数である．

(2) 内容距離による新鮮度

ページ A とページ B 間の違いを表す内容距離は次式で与えられる．

$$dis(A,B) = 1 - sim(A,B) \tag{2}$$

$$sim(A,B) = \frac{v(A)\cdot v(B)}{|v(A)||v(B)|}$$

ここで，$v(A)$，$v(B)$ は，それぞれページ A，B の特徴ベクトルである．

内容距離はページ A がページ B に対して，どれほどの新たな情報が追加されたかということを表している．そこで，ページ A と類似ページとの平均内容距離が大きいほど，ページ A の新鮮度は高いといえる．

☆ 馬らは，新鮮度・流行度を計算するための範囲を時系列文書のさかのぼる遡及範囲と呼んでいる．

内容距離による新鮮度は次のように与える．

$$fresh_{cd}(A,\omega)=\log\left(\frac{1}{m}\sum_{i=1,b_i\in\omega}^{m}dis(A,b_i)\right) \quad (3)$$

ここで，ω は比較範囲内における，ページ A の類似ページ b_i の集合である．

(3) 類似ページの密度による新鮮度

比較範囲内における類似ページの密度 m/n が小さいほど，ページ A の価値は高いと考えられる．そこで類似ページの密度による新鮮度を次のように与える．

$$fresh_{de}(A)=\log_2\frac{n}{m} \quad (4)$$

(4) 時間距離による新鮮度

類似ページとの時間距離が大きいほど，新鮮度が高いと考えられる．そこで，ページ A と類似ページの平均時間距離が大きいほど，ページ A が属しているトピックに新しい動向があったと考えられ，ページ A の新鮮度は高い．ページ A の時間距離による新鮮度を次のように与える．

$$fresh_{td}(A,\omega)=\log\left(\frac{1}{m}\sum_{i=1,b_i\in\omega}^{m}(t(A)-t(b_i))\right) \quad (5)$$

ここで，$t(A)$ はページ A の更新時間，ω は比較範囲内における，ページ A の類似ページ b_i の集合である．

また，最終的な統合新鮮度 $fresh_{\Omega}(A)$ は以下で与えられる．

$$fresh_{\Omega}(A)=\alpha*fresh_{num}(A)+\beta*fresh_{cd}(A,\omega)+\gamma*fresh_{de}(A)+\sigma*fresh_{td}(A,\omega) \quad (6)$$

$$\alpha+\beta+\gamma+\sigma=1.0,$$

$$\alpha\geq 0,\beta\geq 0,\gamma\geq 0,\sigma\geq 0$$

ここで，α，β，γ，σ はそれぞれ重みであり，また ω は比較範囲 Ω 内における，ページ A に類似するページの集合である．

流 行 度

ユーザが興味あるトピックにおいては，内容が類似する変更であっても価値ある変更である．あるページ A が追加，変更された場合，流行度は

(1) 比較範囲内の A の類似ページの密度

(2) A と類似ページとの時間距離

によって決定される．つまり比較範囲内で A に類似するページが多数存在し，それらとの時間距離が小さければ，A の流行度は高いとする．

そこで比較範囲 Ω におけるページ A の流行度を次のように定義する．

$$pop_{\Omega}(A)=e^{\lambda_1 d}+e^{-\lambda_2 t_d} \quad (7)$$

ここで，$\lambda_1(>0)$，$\lambda_2(>0)$ は重み，d は類似ページの密度，t_d は A と類似ページとの平均時間距離である．

3.2.2 更 新 頻 度

毎日更新されるページの更新と，ほとんど更新されないページが更新された場合とでは，ユーザが受ける印象には違いがある場合がある．

ここでは各ページの更新間隔に基づいてサイトの変更価値を評価する．新鮮度を考慮した場合，ページの更新間隔が大きいほど，変更の価値は高いと考える．逆に，流行度を考慮した場合は，頻繁に更新されるページほど，変更の価値は高いと考える．

新鮮度を考慮した場合，更新間隔に基づく評価値 $V_{uf-fresh}(P)$ は次のように与える．

$$V_{uf-fresh}(P)=log(interval(P,n)) \quad (8)$$

$interval(P,n)$ はページ P の平均更新間隔である．

$$interval(P,n)=\frac{t(n)-t(n-1)+interval(P,n-1)\cdot(n-1)}{n}$$

$t(n)$：追加，変更されたページの n 回目の更新時間

ただし，n は更新の回数を表し，$interval(P,n-1)$ はページ P の $(n-1)$ 回目，つまり前回の更新の時点での平均更新間隔である．

また，流行度を考慮した場合，評価値 $V_{uf-pop}(P)$ は次式で与える．

$$V_{uf-pop}(P)=\frac{1}{V_{uf-fresh}(P)} \quad (9)$$

3.2.3 アクセス頻度

ユーザのアクセス頻度が，ユーザの興味を反映していると仮定すると，アクセス頻度が高いページ，トピックへの変更は，その価値が高いと考えられる．アクセス頻度に基づく評価値 $V_{access}(P)$ を次のように与える．

$$V_{access}(P)=\log(\alpha_{t_i}) \quad (10)$$

ここで，α_{t_i} は変更ページ P に対する，一定期間 t_i でのアクセス数とする．

3.3 実験と評価・考察

変更内容の価値推測の有効性を評価するために，日刊スポーツの Web サイト[17] を題材とし，新鮮度/流行度を計算する実験を行った．Web サイトを監視する間隔を約 1 日とし，約 2 日分の変更データ（6401 ページで形成されるサイトに対し，新しく追加された 303 ページ）を対象に計算を行った．題材とした Web サイトのディレクトリ構造は細かく細分化されており，また，ほとんどの追加ページに対し，比較対象となる

表 2 実験結果
Table 2 Experimental result.

	新鮮度	流行度	類似/比較ページ数
平均値	0.450	0.433	0.726
最小値	0.134	0	–
最大値	1	0.894	–
再現率	0.803	0.351	–
適合率	0.564	0.540	–

ページは 1 階層下までに格納されている．そのため今回は比較範囲の選択を 1 階層下のディレクトリ内までとした．事前にパラメータ設定のための予備実験を行い，その結果，類似ページ判定の閾値は 0.6，新鮮度（式 (6)）に関する重み付けパラメータ α，β，γ，σ は，それぞれ 0.4，0.4，0.1，0.1 と設定した．実験の結果を**表 2** に示す．比較ページが存在しない追加ページに関しては，各種新鮮度，類似ページの密度を 0，類似ページとの時間距離を 1 とした．**図 6**，**図 7** は新鮮度，流行度の数値のばらつきを示している．

図 6 から，新鮮度に関しては 0.25 前後にまとまった数値が得られた．また，新鮮度 1 のページが 71 ページあるが，これは比較対象となるページが存在しないページである．流行度に関しては多少ばらつきがあるが，この要因として，流行度を計算する際に用いている類似ページとの時間距離が大きく影響したと考えられる．今回題材とした Web サイトはいわゆるニュースサイトであり，過去のニュースページも多くアーカイブしてあるため，そうしたページとの時間距離に大きくばらつきが生じたものと思われる．これに対し，新鮮度計算では時間距離に関する重み付けパラメータを小さく設定してあったため，その影響が小さかったと考えられる．

新鮮度に関する再現率，適合率は 0.803，0.564 であった．新鮮度が 0.25 以上の変更ページをシステムの解とし，ユーザによって正解ページを選んでいる．また，新鮮度が 1，すなわち比較対象となるページが存在しなかったページを除いて計算した場合，再現率は 0.718，適合率は 0.650 であった．これらの結果に対し，今回用いた各 Web ページの構成による影響を考えることができる．今回用いた Web ページのほとんどは，baseball，soccer などのサイト内の各トピックへのアンカーが存在していた．つまり，ほとんどのページに共通の情報として各トピックへのアンカーが存在し，ページの本文となる部分の内容が少ないページでは，本文にあたる内容が新鮮であっても，その他の共通する部分により結果として類似と判断されたと考えることができる．これには各 Web ページの構造解析などによって，内部リンクのアンカーテキストは無視するなどの改善が考えられる．

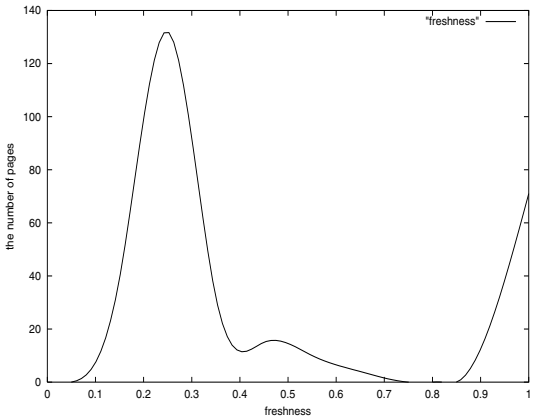

図 6 新鮮度
Fig. 6 Distribution of freshness.

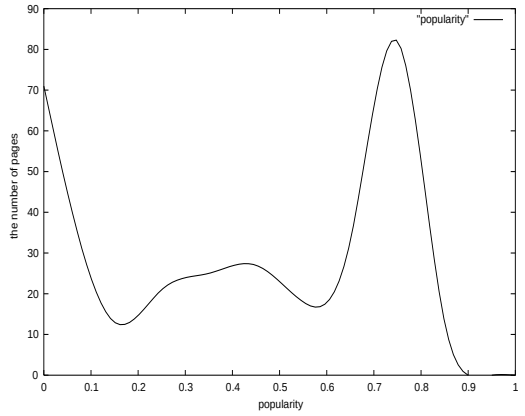

図 7 流行度
Fig. 7 Distribution of popularity.

流行度に関する再現率，適合率は 0.351，0.540 であった．流行度が 0.75 以上の変更ページをシステムの解とした．また，流行度が 0，すなわち比較対象となるページが存在しなかったページを除いて計算した場合，再現率は 0.753，適合率は 0.430 であった．これには，先に述べた時間距離のばらつきによる影響が大きいと考えられ，対象とする Web サイトにあわせて，時間距離の上限を設定するなどの工夫が必要であると思われる．

これらの結果より，ある程度の精度を保ったうえで，数多い変更の中からユーザへ通知する必要あるページを選択することが可能であると考える．また，比較対象が存在しない場合を除いた再現率と適合率も考慮すると，新鮮度・流行度の計算は，構造的な比較範囲に

対しても有効であると考えることができる．

また，今回の実験の題材とした Web サイトでは，変更ページと同一ディレクトリ内において類似ページの占める割合が平均しておよそ 72.6%であることから，同一ディレクトリ内のページは同一トピックであるという仮説が成立していると考えることができる．しかしながら Web サイトの性質上，あるディレクトリ（flash）については，その日のニュースを話題を無視して，まとめて格納するためのディレクトリであり，この中にはサイト全体の約 1%のページが格納されている．実験で扱った変更のうちの約 14%がこのディレクトリへの変更であったため，このこともまた誤差が生じる要因の一部であったと思われる．

4. プッシュ型変更通知とパーソナライズ

本章では，WebSCAN のプッシュ型の変更通知機構について述べる．

WebSCAN のプッシュ型変更通知の特徴として，次の 3 つがあげられる．

(1) 配信コンテンツとパーソナライズ機構の分離

Web サイト内のページの追加や更新に対する価値は，ユーザごとに大きく違う．そのため WebSCAN では，変更に関する同一の情報を各ユーザに配信し，クライアント側で独自の通知情報を抽出し，変更情報としてユーザに呈示する．

(2) 動的な変更情報（HTML）の表示

配信される変更データから，クライアント側での各ユーザ・プロファイルを利用し，動的に HTML 文書を作成し，表示する．

(3) 変更ページの内容表示

どのページが変更，追加されたということよりも，どんな内容が変更，追加されたかということの通知が必要である．つまり，変更されたページあるいは部分に対して，その情報の内容に関する通知が望ましい．今回は，ページの内容をユーザが簡潔に把握できるものとして，ページタイトル，見出し，ページ内で利用されている画像などを呈示する．

以下，4.1 節でユーザへ配信されるコンテンツについて，4.2 節でユーザ側での変更情報のパーソナライゼーションについて述べる．

4.1 配信コンテンツ

ユーザに配信するための変更データは，前回の配信時以降に Web サイト内に生じた変更，追加ページに関するデータである．この変更データは，対象とする Web サイトの構造を反映し，また新鮮度/流行度，アクセス頻度による評価値，更新頻度に基づく評価値，ページ URL，タイトルなどの情報が記述される．

4.2 変更情報のパーソナライズ

ユーザ側ではユーザ・プロファイルにより，Web サイトの指定，トピックの指定，変更価値の最終的な計算などを行い，HTML 変換により変更情報の呈示を行う．今回，我々が考えているパーソナライゼーションとして，次の 3 つをあげる．

(a)Web サイトの指定

各ユーザの通知して欲しいサイトの指定が可能である．つまりサーバ側で監視を行っている複数の Web サイトから，自分の要求する Web サイトに関する変更データのみを取り出すことができる．

(b) トピックの指定（優先度）

Web サイト内には，複数のトピックが存在するものである．また各ユーザごとに，このトピックの話題は絶対に伝えて欲しいとか，このトピックに関する情報は必要ないなど，トピックに応じた要望や優先度が存在するはずである．そこで，ユーザの各トピックに関する優先度を，配信される変更データに対して，URL によるディレクトリ指定することで表現する．またトピックの優先度を重みで表現し，それぞれのトピックの変更価値を重み付けする手法も考えられる．

さらに，どの変更を選択するかというフィルタリングだけでなく，レイアウトについての記述もまた可能である．たとえば，あるトピックに関しては，すべてトップに表示するとか，全体の何割はこのトピックに関する情報を表示する，また評価値に応じて色を変えて表示するといったことも考えられる．

(c) 価値計算

最終的な変更価値は，クライアント側で各ユーザごとに決定される．つまり，変更価値を決定するための評価要素として，新鮮度/流行度，アクセス頻度，更新頻度などの情報はサーバ側から変更データとして与えられ，最終的な変更価値はクライアント側のユーザ・プロファイルによって決定される．ある変更ページ A に対する変更価値 $change\_worth(A)$ は次のように定義する．

$$
\begin{aligned}
&change\_worth(A) = \\
&\quad \omega_1 * fresh_{\Omega}(A) + \omega_2 * pop_{\Omega}(A) \\
&\quad +\omega_3 * V_{upd-freq}(A) + \omega_4 * V_{access}(A) \\
&\omega_1 + \omega_2 + \omega_3 + \omega_4 = 1, \\
&\omega_1 \geq 0, \omega_2 \geq 0, \omega_3 \geq 0, \omega_4 \geq 0
\end{aligned}
\tag{11}
$$

$$
V_{upd-freq}(A) = \begin{cases} V_{uf-fresh} & (\omega_1 \geq \omega_2) \\ V_{uf-pop} & (\omega_1 < \omega_2) \end{cases}
$$

ここで，ω_1，ω_2，ω_3，ω_4 は重みである．

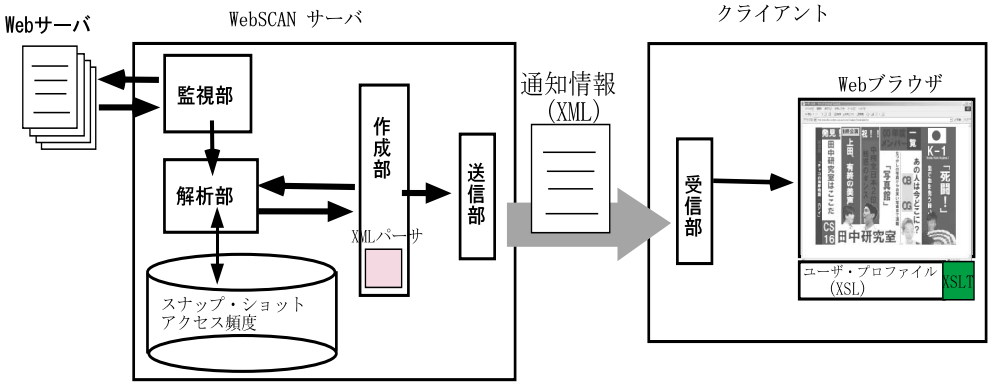

図 8 システムモデル
Fig. 8 System model.

これらの重み付けパラメータはユーザの興味に応じて決定できる．たとえば新鮮度を重視して，変更価値を決定したい場合は，それに関する重み ω_1 を増やせばよい．

また，価値計算手法の 1 つの提案として，新鮮度/流行度とアクセス頻度を統合した計算手法をあげる．これは，アクセス頻度がユーザの各トピックへの興味を反映していると仮定して，アクセス頻度が低いトピックほど新鮮度を重視し，流行度をあまり考慮しないという計算手法である．これを次式で定義する．

$$change\_worth(A) = \begin{cases} \omega_1 * V_{fresh}(A) + \omega_2 * V_{uf-fresh}(A) & (\mu \leq \theta) \\ \omega_1 * V_{pop}(A) + \omega_2 * V_{uf-pop}(A) & (\mu > \theta) \end{cases} \quad (12)$$

$$\omega_1 + \omega_2 = 1, \quad \omega_1 \geq 0,\ \omega_2 \geq 0,$$

$$V_{fresh}(A) = (1-\mu) * fresh_{\Omega}(A) \quad (13)$$

$$V_{pop}(A) = \mu * pop_{\Omega}(A) \quad (14)$$

$$1 \geq \mu \geq 0$$

つまり，ユーザのアクセス頻度 μ がある閾値 θ 以下である場合は新鮮度を重視，そうでない場合は流行度を重視する．この手法は，ユーザが新鮮度，流行度のパラメータ設定を行わず，システムがパラメータを自動的に設定する必要がある場合に有効であると考えられる．

5. プロトタイプシステム：WebSCAN

WebSCAN のプロトタイプシステムを現在開発中である．**図 8** にそのシステムモデルを示す．今回のプロトタイプシステムでは配信する変更データとして

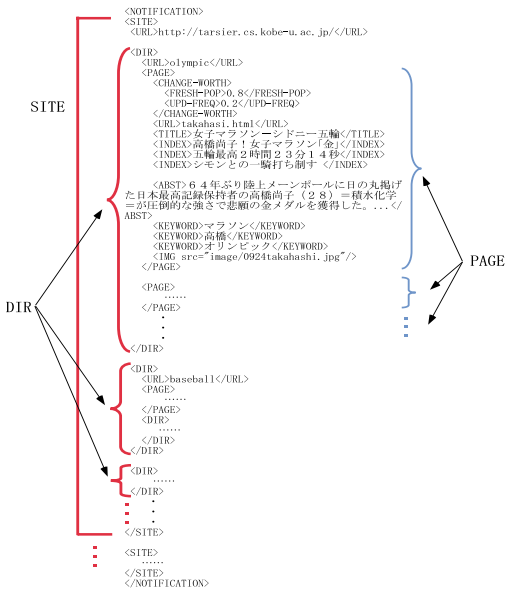

図 9 XML データの一例
Fig. 9 Sample of XML data.

XML [18] を利用し，ユーザ・プロファイルを XSL [19] を用いて表現する．

5.1 システム構成

システム構成として，サーバ側は監視部，解析部，作成部からなる．監視部では，サイト内の全ページをタイムスタンプ，サイズによって監視し，変更を検知した場合，解析部で変更解析を行う．その後，作成部にて，配信用の XML ドキュメントを作成する．XML ドキュメントの例を**図 9** に示す．クライアント側は，受信部と変更情報を表示するための Web ブラウザからなる．ブラウザでは XSL ファイルによって，最終的な価値計算，好みのトピックなどを考慮して変更情

```
<?xml version="1.0" encoding="Shift_JIS"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/TR/WD-xsl">
<xsl:template match="/">
    ......
<xsl:if test="NOTIFICATION/SITE[URL='http://tarsier.cs.kobe-u.ac.jp/']">
<xsl:apply-templates select="//DIR[URL='olympic']"/>
<xsl:apply-templates select="//DIR[URL='baseball']"/>
</xsl:if>
    ......
</xsl:template>

<xsl:template match="//DIR[URL='olympic']">
<tr>
<td width="74%" height="120" bgcolor="#D67030"
  style="font-size: 1    ……
    ......
</xsl:template>

<xsl:template match="DIR[URL='baseball']">
    ......
</xsl:template>

<xsl:script language="JavaScript"><![CDATA[
  w1=0.5; w2=0.1;
  total=0;top=0;
    ......
 function totalworth(p){
   freshpop=p.selectNodes("//CHANGE-WORTH/FRESH-POP");
     ......
     total = w1*v.nodeTypedValue+w2*u.nodeTypedValue;
   return(total);
 }
    ..........
]]></xsl:script>

</xsl:stylesheet>
```

①

②

トピック'olympic'に関する処理

トピック'baseball'に関する処理

スクリプトの記述

図 10 XSL ファイルの一例
Fig. 10 Sample of XSL data.

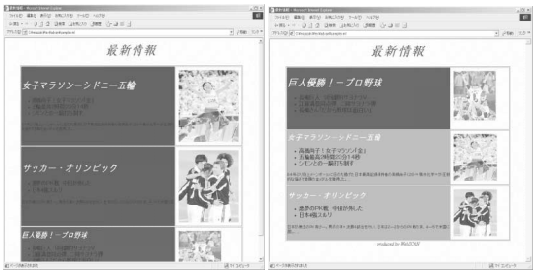

図 11 ユーザ・ビューの違い
Fig. 11 Comparison of user views.

報が選択，表示される．XSL の一例を**図 10** に示す．

5.2 ユーザ・プロファイルの表現

図 10 の XSL ファイルの例では，① において対象とするサイト（http://tarsier.cs.kobe-u.ac.jp）を指定している．

また，XSLT の順序処理を利用して，② において，トピックの指定を行っている．この例では'olympic'，'baseball' の 2 つのトピック内の変更だけを抽出，表示する．

また，このように，内部にスクリプトを用いることで，変更価値の再計算が可能となり，それに応じた表示方法の指定が可能となる．

XSL ファイルを用いたユーザ・プロファイルにより，**図 11** のように，同じ変更データ（XML データ）から，各ユーザごとに独自の変更情報が呈示される．これは各ユーザ独自のプロファイルにより，トピックへの優先度の指定，最終的な変更価値の決定方法，レイアウトの指定などが異なるためである．

現在のプロトタイプシステムは，Perl により Windows2000 上で実装されている．サーバ上で 1 つの Web サイトを対象に監視する監視部，サイト内の変更に対して，新鮮度，流行度，更新頻度などの各パラメータの計算を行う解析部が稼動する．

また，クライアント側の変更通知表示には Web ブラウザを利用している．

6. 関連研究

Web ページの変更発見に関する研究が C3 project[9),10)] で行われている．C3 ではユーザが変更監視対象を指定すると，データベースへの問合せを行う．C3 の特徴として，2 つの構造データ間の変更を呈示し，また変更の検出の際にはデータ構造を考慮する．一方，WebSCAN ではコンテンツと構造に基づいて変更の評価を行う．

Netmind[11)] は Web サーチエンジンを拡張した，URL の変更監視システムである．Netmind は電子メールにより Web ページの更新を通知するものであるが，変更を解析，評価するといったことは行っていない．

WebGUIDE[12)] は再帰的な文書比較による Web ページと Web の構造に対する変更追跡システムである．WebGUIDE は Web ページの更新を追跡，表示する AIDE[13)] と，ユーザへの視覚的なナビゲータとしての Ciao[14)] の 2 つから構成されている．Ciao[14)] ではユーザは問合せとブラウジングが行える．WebGUIDE の特徴としては，再帰的な文書比較と視覚的ナビゲータによる違いの視覚化であるが，他のシステムと同様に変更の意味的なものを考慮するものではない．

WebCQ[15),16)] は Web ページ内の変化を発見し，ユーザにパーソナライズされた通知を行うシステムである．WebCQ の特徴は，監視・追跡する変更の種類の豊富さと，ユーザごとのパーソナライズされた変更情報の呈示である．しかしながら，他の既存システムと同様に，変更の意味的な評価は行っておらず，また新たに作られたページを通知することもない．また，変更通知はユーザの興味に基づいて行われるため，ユーザの興味は明確なものである必要がある．Web 上の新しく作られる情報は未知で，そのタイミングも決まっていないことを考えると，ユーザのプロファイルを特定させるのは簡単ではない．我々の研究のアプローチでは新情報は等しく新鮮ではなく，変更に対し意味的な評価を行う．さらに WebCQ ではユーザごとの変更情報を作成し，ユーザに通知するが，我々の研究では同一の変更情報をユーザに配信し，ユーザ側でパーソナライズすることを目指している．

7. おわりに

本論文では Web サイトに対する変更監視・解析・通知システム—WebSCAN—について提案した．これはプッシュ型配信機構に基づいた，ユーザに価値ある変更通知を行うための機構である．

Web の変更の時系列的な特徴を考慮し，変更の意味的な評価として，新鮮度/流行度という尺度を導入し，変更内容と Web サイトの構造に基づいた変更解析について述べた．また変更通知の手段として，プッシュ型配信機構による変更情報の配信とユーザ・プロファイルによるパーソナライゼーションについて述べた．

また，実験によって提案手法の検証を行った．実験結果によって，以下のようなことが分かった：

- **変更価値などに基づいてユーザへ重要な変更だけを通知することが可能である**．
  本論文で提案した手法は，一定の精度を保ったうえで，たくさんの重要な変更を検出することが可能である．
- **時間のみではなく，構造を考慮して情報の新鮮度・流行度を計算することが可能である**．
  我々は，馬らが定義した時系列データの特徴量（新鮮度と流行度）の時間による計算範囲を Web サイトの構造による範囲へ拡張し，それに基づいて変更価値の計算を行う手法を提案している．実験結果により，この拡張は有効であると考える．
- **Web ページ（サイト）の構造に基づく比較範囲の選択が有用である**．
  特に，URL パスによって構成された Web サイトの木構造では，同じディレクトリに格納されているページは同じトピックに属するという仮説は有効であると考える．

今後の課題として，その他の解析要素，たとえばリンクの変更，追加などの扱い，また WebSCAN の実装の改良をあげる．また，複数のサイトを取り扱った実験，検証も必要である．

さらに，フィードバック機構を用いた，動的なユーザ・プロファイルの更新もまた今後の課題と考えている．

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 12680416）の援助を受けており，また，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）によっております．ここに記して謝意を表すものとします．

参考文献

1) 馬 強，角谷和俊，田中克己：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，情報処理学会論文誌データベース，Vol.41, No.SIG6 (TOD7), pp.46–57 (2000).
2) 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.SIG8 (TOD4), pp.141–157 (1999).
3) Tajima, K., Mizuuchi, Y., Kitagawa, M. and Tanaka, K.: *Cut as a Querying Unit for WWW*, Netnews, e-mail. Hypertext, pp.235–244 (1998).
4) Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J.: Graph structure in the web, *Proc. WWW9* (2000).
5) Pu, C.: Infosphere: Smart Delivery of Fresh Information, ACM SIGMOD 日本支部第 15 回大会，pp.1–18 (2000).
6) WebBase. http://www-diglib.stanford.edu/testbed/doc2/WebBase/
7) Cho, J. and Garcia-Molina, H.: Synchronizing a Database to Improve Freshness, *Proc. ACM SIGMOD International Conference*, pp.117–128 (2000).
8) 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，アドバンスド・データベース・シンポジウム'99（ADBS'99），pp.141–150 (1999).
9) C3 Project. http:// www.db.stanford.edu/c3/c3.html.
10) Chawathe, S.S. and Garcia-Molina, H.: Meaningful change detection in structured data, *Proc. SIGMOD'97*, pp.26–37 (1997).
11) NetMind. http://www.netmind.com/
12) Douglis, F., Ball, T., Chen, Y-F. and Koutsofios, E.: WebGUIDE: Querying and navigating changes in web repositories, *Proc. WWW5*, pp.1335–1344 (1996).
13) Douglis, F. and Ball, T.: Tracking and viewing changes on the web, *Proc. USENIX Technical Conference*, pp.165–176 (1996).
14) Chen, Y-F., Fowler, G.S., Koutsofios, E. and Wallach, R.S.: Ciao: A graphical navigator for software and document repositories, *Proc. International Conference on Software Maintenance*, pp.66–75 (1995).
15) WebCQ. http://www.cc.gatech.edu/projects/disl/webcq/
16) Liu, L., Pu, C. and Tang, W.: WebCQ-detecting and delivering information change on

the Web, *Proc. CIKM'00* (2000).
17) 日刊スポーツ. http://www.nikkansports.co.jp/
18) W3 consortium. Extensible Markup Language (XML)1.0. http://www.w3.org/TR/REC-xml/
19) W3 consortium. Extensible Stylesheet Language (XSL)1.0. http://www.w3.org/TR/xsl/

(平成 12 年 12 月 20 日受付)
(平成 13 年 2 月 28 日採録)

（担当編集委員　牧之内 顕文）

宮崎　慎也（学生会員）
2000 年神戸大学工学部情報知能工学科卒業，現在，同大学院自然科学研究科修士課程在学中．データベース，放送型情報メディアに興味を持つ．

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．IEEE Computer Society 学生会員．

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（情報メディア科学専攻）専任教授，2001 年京都大学大学院情報学研究科（社会情報学専攻）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 各会員．