データベースシステム　122－68
(2000. 7. 28)

Webサイトの変更監視・解析と放送型変更通知機構

宮崎慎也[†]　馬　強[††]　田中克己[††]

本稿ではWebサイトを対象に，その中のページの追加，更新を監視し，サイト内のページの最新の情報をユーザに配信する機構について述べる．すべての変更を伝えるのではなく，その変更が価値ある変更であるか，ユーザに対してどの変更を伝えるべきかということを判断するために，追加，更新されたページに対してその価値を非類似度，変更量，アクセス数，更新頻度という要素に基づいて評価し，その評価を反映した変更情報を作成する．また作成した変更情報は放送型配信を用いてユーザに効果的な変更通知を行うものである．

A Monitoring and Change Notification System for Web Sites

SHINYA MIYAZAKI,[†] MA QIANG[††] and KATSUMI TANAKA[††]

In this paper, we propose a monitoring and change notification system for Web sites. The changes of Web sites are monitored and analyzed with the contents, update time and access history. Based on the analysis, the notification information is generated and sent to user automatically.

1. はじめに

WWW(World Wide Web)の利用者には，定期的あるいは非定期的にアクセスするページやサイトがある．ユーザがそのようなサイトにアクセスした際に，前回のアクセスから変更された部分がない，あるいは変更があったとしてもユーザがそれに気づかないといった場合があるとが考えられる．これはユーザにとっては無駄足でしかないし，一方ページの製作者にとってはせっかくアクセスしてくれたユーザに新たな情報を伝えることができない場合がある．たとえばニュースサイトでは最新のニュースにどんなものがあるか知りたい，また企業のサイトで商品の紹介のようなページでは新商品をユーザに見てもらいたいなど，サイト内の変更を効果的にユーザに通知するようなサービスが望まれる．

ユーザ側からWebサイト内のページの追加や更新を自動的に知るためのアプローチとして，いわゆる自動巡回ソフトと呼ばれるようなアプリケーションの更新調査の機能の利用が考えられる．しかしながら，これらのアプリケーションには次のような問題が挙げられる．

- ファイルのタイムスタンプやサイズのみの判断である．すべての変更は同様に扱われる．
- 変更の有無しか伝えないものが多い．どういった変更かわからない．

また，サイト側，ページの作成者側からのアプローチとしては，バナー広告で宣伝をしたり，メールマガジンの配送などが考えられるが，

- 確実にその広告がユーザの目に触れるとは限らない．
- 人手による作業では，大きな労力を必要とする場合がある．

など問題点がある．

本研究では，各Webサイトの変更をコンテンツの変更量，アクセス数，更新間隔と非類似度に基づいて評価し，ユーザにプッシュ技術を用いて変更通知を行う手法を提案する．本研究で提案する手法として，次のような特徴が挙げられる．

- **Webサイト全体を対象とした変更監視・解析**
  一つのページだけを監視するのではなく，サイト全体を監視する．つまりWebサイトを評価の対象としている．
- **Webサイトの追加，変更に対する内容の評価**

† 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

変更前後のWebサイトのコンテンツの非類似度に基づいて，Webサイトの変更を評価している．

- プッシュ技術を用いた変更通知
  サイト変更の評価に応じて，ユーザへの通知情報を生成．プッシュ技術を用いて生成された変更情報をユーザに配信する．[1)]
- ユーザに呈示する変更情報として，HTMLページの作成
  ユーザに呈示する変更情報として，評価された価値を反映した変更情報をHTMLで自動生成する．評価の高い変更ほど，大きく，目立つよう表現する．

Webサイトの変更の監視，評価と，評価に基づく変更情報の呈示により，その時点でのサイト内の最新情報はどれか，どのページが価値あるページかという認識をユーザに可能にさせる．

以下，2章で関連研究について述べる．3章で変更の監視，解析について述べる．4章で放送型変更通知について述べる．また5章で試作のプロトタイプシステムについて述べる．最後に6章でまとめと今後の課題について述べる．

2. 関連研究

WebCQ[2)]はジョージア工科大のCalton Puらによるプロジェクトで，Webページを監視し，その変更をユーザに通知するサービスである．ユーザは監視するWebページのURLと，変更の種類（キーワード，リンク，画像など全9種類），また通知の間隔（一日一回など）を登録し，登録した種類の変更のみが通知される．通知の方法は現在はメールによる通知のみで，メールを受け取ったユーザは，変更情報が記述されているWebページにアクセスし，変更の詳細を知ることができる．しかしながら，対象は既存のWebページに対する変更のみで，サイト内の新しいページの追加は知ることができない．また発生した変更を忠実に通知しているが，変更された内容が価値あるものかどうか評価するということなどは行っていない．

Site Outlining[3)4)]はIBMの武田らによって研究，開発されているシステムで，ユーザのWebサイトからの情報収集，検索を支援するための技術である．Webサイトを動的な情報源として捕らえ，テキスト情報と，そこから自然言語処理，情報抽出によってメタ情報を取得し，その差分から時系列変化の内容を把握し，またそれらを複数のサイト間で比較することで，サイトのアウトライン情報を呈示するものである．しかしながらこれらはユーザの情報収集の意思決定支援のためのもので，やはり情報の価値を評価するなどは行っていない．本研究では，情報の更新間隔や量などの外観的な側面で情報を解析するが，加えてその類似性を考慮し，各情報に価値を付与し，そこからサイト内の最新の情報を積極的に呈示しようとするものである．

また，ユーザ側から自動的にページの変更を知る手段として，変更通知機能を有するWebの自動巡回アプリケーションやWebページの更新チェックを行うソフトウェアが挙げられる．例えば，フリーのソフトウェアとして普及しているWWWC[5)]というWeb巡回ツールでは，HTMLのMETAタグから更新を検知する機能もあるが，基本的に指定Webページのタイムスタンプとファイルサイズによって更新を判断する．しかしながらたいていの場合，これらの変更通知は変更の有無を伝えるだけのもので，変更の解析などは行わなず，またユーザが知ることができるのは指定Webページの変更の有無だけで，その概要を把握することはできない．

馬らは，データの時系列性を考慮した情報フィルタリング[6)]について提案している．これは配信されるニュース記事の時系列的な特徴量として新鮮度，流行度，緊急度を定義し，これ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の提案である．本研究でも変更内容の新鮮度というものを考慮するが，対象が配信されるニュース記事とWebサイト内のページであるという違いがある．

また，これらの研究は基本的にページ単位での変更監視，追跡であり，新しくページが作成された場合などに対応しない．本研究ではWebサイト全体を対象としてページの追加，更新を監視し，それを積極的にユーザに通知しようとするものである．そのため，Webサイトを閲覧する一般ユーザにはもちろん，サイトの内部ユーザ（例えば企業ならその社員，組織・団体などではそこに属する人々）に対しても有用なものであると考える．

3. Webサイトの変更監視・解析

Webサイト内の各ページに対し監視を行い，変更，あるいは追加されたページを抽出する．サイトの変更価値を追加，変更されたページの非類似度，変更量，アクセス数と更新頻度に基づいて評価する．図1にそのモデルを示す．3.1節では，評価モデルについて述べる．3.2節ではページ内の変更について述べる．3.3

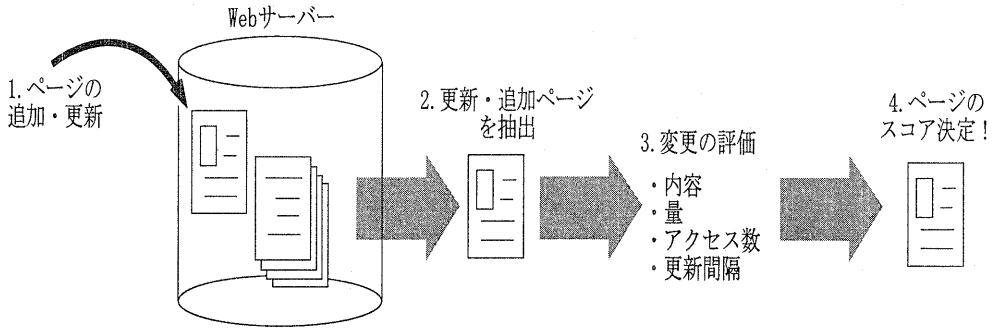

図 1　Web サイトの監視・解析モデル

節ではページの追加について述べる．

3.1 解析要素

3.1.1 非類似度

追加，変更された内容*の価値は各ユーザによって異なる．

ここでは，内容の非類似度に基づいてサイトの変更価値を評価する．つまり，その変更はどれほどユーザに伝えるべき内容であるかを評価する．サイト内の既存の各ページと比較して，追加又は更新された内容の非類似度が高いということは，そのページの内容が今までの内容にはなかったような話題，事柄についてのものであると判断でき，それはユーザにとって新鮮であり，かつ価値があると判断する．

追加・変更部分の特徴ベクトル $\vec{v}_1$ と比較の対象の特徴ベクトルを平均化したベクトル $\vec{v}_2$ との非類似度 $dissim(\vec{v}_1, \vec{v}_2)$ を，特徴ベクトル間の類似度 $sim(\vec{v}_1, \vec{v}_2)$ を用いて次式として与える．

$$dissim(\vec{v}_1, \vec{v}_2) = (1 - sim(\vec{v}_1, \vec{v}_2)) \quad (1)$$

$$sim(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1 \cdot \vec{v}_2}{|\vec{v}_1||\vec{v}_2|} \quad (2)$$

ここでは対象とする二つのベクトルに対する非類似度を，特徴ベクトル間のコサイン相関値による類似度から求めている．これによって算出された値をその変更の非類似度での評価値とする．ただし，比較対象は，新規ページ追加ならば他の既存ページ，既存ページ内の変更の場合は同一ページの他の部分である．

非類似度が高ければ，変更価値が高いと考え，非類似度による変更価値 $V_{dissim}(P)$ は次のように与える．

$$V_{dissim}(P) = e^{dissim(\vec{p},\vec{q})} \quad (3)$$

$\vec{p}$: 追加，変更部分の特徴ベクトル

$\vec{q}$: 比較対象とする部分の特徴ベクトル

3.1.2 変更量

追加もしくは更新されたコンテンツの量も，サイトの変更価値に影響を与える．たとえば数行だけ追加された場合よりも，数十行の追加された場合の方が一般的には情報量が多いと考えられる．

ここでは追加，変更部分のコンテンツの量，つまり追加，変更された部分の単語数に基づいて変更価値を評価する．

追加，変更されたコンテンツの差が大きければ，変更の価値が高いと考え，追加・変更されたページの変更量に基づく評価値 $V_{quantity}(P)$ は，次式で与える．

$$V_{quantity}(P) = \log C \quad (4)$$

ただし，C は追加されたページ，もしくは変更されたページの変更部分から HTML のタグ，コメントを除去した残り単語数とする．

3.1.3 アクセス数

アクセス数が多いページが変更された場合，一般的に価値の高い変更であると考えることができる．

アクセス数におけるページ変更の評価値 $V_{access}(P)$ は次式として与える．

$$V_{access}(P) = \log a_{t_i} \quad (5)$$

ただし，a_{ti} は更新された各ページの，期間 t_i におけるアクセス数である．

既存ページへの変更に対しては，そのページのある時間幅でのアクセス数に基づいてサイトの変更価値を評価する．ページが新しく作成された場合は，当然アクセス数を考えることはできないので，他の各既存ページのアクセス数のうち，最高のページと同等とする．

*本論文では，サイトにおける内容の追加，変更は，基本的にテキストの追加，変更として考えている．

3.1.4 更新頻度

更新間隔が長いページの変更価値は高いと考えられる．たとえば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．各ページの平均更新間隔を評価値として考える．ただし，この場合も新しく作成されたページに関しては，更新頻度を得ることはできないため，他のページの評価値の中で最高値と同等とする．

更新間隔に基づく評価値 $V_{interval}(P)$ は次のように与える．

$$V_{interval}(P) = \frac{t(n) - t(n-1) + V_{interval}(P, n-1) \cdot (n-1)}{n} \quad (6)$$

$t(n)$: 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，$V_{interval}(P, n-1)$ はページ P の前回の更新の時点での平均更新間隔である．$V_{interval}(P)$ はページの平均更新間隔として与えられる．

これらの各要素からそれぞれの評価値を算出し，各変更ページに対し総合の変更価値 $Score(P)$ を与える．対象とするページあるいはページ群 P の上に述べた各評価要素での評価値をそれぞれ，$V_{dissim}(P)$，$V_{quantity}(P)$，$V_{access}(P)$，$V_{interval}(P)$ とすると，P の追加・変更に対する変更価値 $Score\ (P)$ は次式で与える．

$$Score\ (P) = w_1 \cdot V_{dissim}(P) + w_2 \cdot V_{quantity}(P) + w_3 \cdot V_{access}(P) + w_4 \cdot V_{interval}(P) \quad (7)$$

w_1, w_2, w_3, w_4: 重み

ここで，各評価値に対する重み w_1, w_2, w_3, w_4 によって，どの要素を重視して評価するかが決定される．この重み w_n はユーザの意思を反映して決定される．例えば，アクセス数の多いページの変更を重要視したければその重みを，またとにかく今までとは違う内容の変更を重視したければ非類似度の重みを調整すればよい．

3.2 ページ内の変更

既存のページが更新される場合（図 2）について，特徴的なものとして，新しい情報が次々に追加されていく場合がある．これは Web 日記や Web 掲示板ページなどでみられる．また，基本的に 1 つのページ内の内容は，あるテーマや話題に沿った，同一の方向性を持った内容であると考えられる．そこで，ページ内への情報の追加，変更には同一ページ内で内容を比較することを考える．

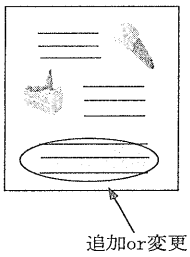

図 2 既存ページ内の変更

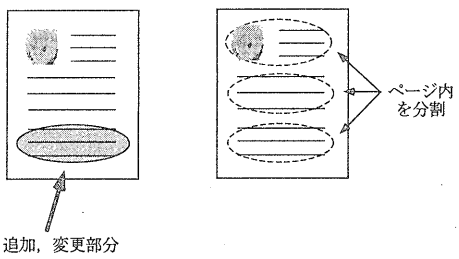

図 3 ページ内の変更の例

ここでは次のようにして情報の非類似度に基づく評価を行う．(図 3)

(1) 追加・変更された部分を更新前ページとの差分から特定する．

(2) 特定した変更部分を囲む HTML の最小のタグペアを抽出する．

(3) ページ全体を（2）で抽出したタグペアと同階層の深さに切り分ける．

(4) 分割した各部分との非類似度を算出する．

また，変更量は変更前後のページの差分から特定し，アクセス数はそのページのアクセス数を取得する．更新頻度は前回の更新記録から算出する．各要素での評価値から，(7) 式に基づいて，変更されたページ p の変更価値は次で与えられる．

$$Score(p) = w_1 \cdot V_{dissim}(p) + w_2 \cdot V_{quantity}(p) + w_3 \cdot V_{access}(p) + w_4 \cdot V_{interval}(p)$$

3.3 新規ページ追加

サイト内へのページの追加は，単一ページの追加と複数ページの追加の二種類があると考えられる．ここでは，それぞれの場合に分けてその解析について述べる．

3.3.1 単一ページ追加

新しくページが追加された場合について，内容の非

類似性を求める際に対象とする範囲としてサイト内の全ページを対象とするのは問題がある．つまり，比較する対象範囲を決定する必要がある．たとえばニュースサイトでは政治，経済，社会などのように，それぞれのジャンルに基づいて各ページが存在している．政治のニュースに関するページと，経済ニュースに関するページとではその内容が似てないのは当然である．政治のニュースは，同じ政治のニュースと比較されることに意味があると考える．

ここでは，同じ URLpath を持つ各ページは，内容的に類似した集まりである可能性が高いと考える．つまり，あるページ P に対して，たとえば P の URL が

http://www.aaa.bbb.ccc/xxx/P

で表されるとき，P の URLpath

http://www.aaa.bbb.ccc/xxx/

を含んでいるページは，P と同一ジャンルの内容である可能性が高いと考える．

そこで，あるページ P の追加を検知した場合，P の URLpath と同じ URLpath に所属するページ，あるいは P の URLpath を含む URL を持つページを一つのカテゴリとして考え，そこに属する各ページに対して P の非類似度に基づく評価を行う．

また追加ページの平均更新間隔などは他のページのうち最高値のものと同等とし，また変更量はそのページの文字数をカウントする．またアクセス数は他のページのうち最高のものと同等とし，各要素に基づく評価値を算出する．

各要素での評価値から，(7) 式に基づいて，追加されたページ p の変更価値は

$$Score(p) = w_1 \cdot V_{dissim}(p) + w_2 \cdot V_{quantity}(p) + w_3 \cdot V_{access}(p) + w_4 \cdot V_{interval}(p)$$

で与えられる．

3.3.2 複数ページ追加

新しく Web ページが作成される場合，複数のページがひとつのグループとしてまとめて追加される場合が考えられる． 図 4 は複数のページがまとめて追加される例で，各ページの URL からサイトを木構造で表現したものである．末端のノードが各 Web ページである．

このように複数ページがまとめて追加された場合は，それらをひとつのグループとして扱い，そのグループをひとつの単位として解析するべきである．

各評価値の算出は，このグループを構成するページ群を 1 つのページのように見立てて算出する．

この場合，非類似度に基づく評価での比較対象範囲

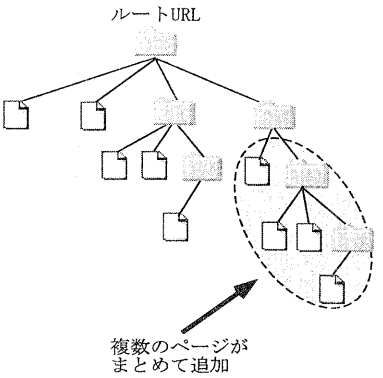

図 4 複数ページの追加

として，次のように考える．(図 5)

(1) サイト内の各ページに対して，その URL から木構造を抽出する．
(2) 追加された各ページを含む最小の部分木を抽出．
(3) (2) で抽出した部分木と，共通の親を持つ各部分木を抽出．得られない場合は，一階層上で抽出を試みる．
(4) 抽出された各グループ間で非類似度の算出を行う．

また，変更量としては追加部分全体での量を考え，更新間隔，アクセス数に基づく評価値は既存ページ中の最高値と同等とする．各要素での評価値から，(7) 式に基づいて，追加ページ群 G の変更価値は，

$$Score(G) = w_1 \cdot V_{dissim}(G) + w_2 \cdot V_{quantity}(G) + w_3 \cdot V_{access}(G) + w_4 \cdot V_{interval}(G)$$

で与えられる．

4. 放送型変更通知

ユーザに対しての通知には効果的な情報呈示方式が重要である．ユーザへに呈示する変更情報には，各ページあるいはグループの評価された価値が反映される必要がある．評価の高いページほどユーザに的確に伝える必要があり，またユーザが意識せずに各ページの評価を認識できることが望ましい．

ここで我々は前章で述べたサイトの変更価値関数 (式 (7)) に基づいて，効果的な情報呈示手法を提案する．前述の方法で各追加・変更ページの評価を行い，サイト内のページのランキングする．そしてその評価に基づいて，評価の高いページの内容ほど，大きく，目立つようにユーザへの変更呈示ページをレイアウト

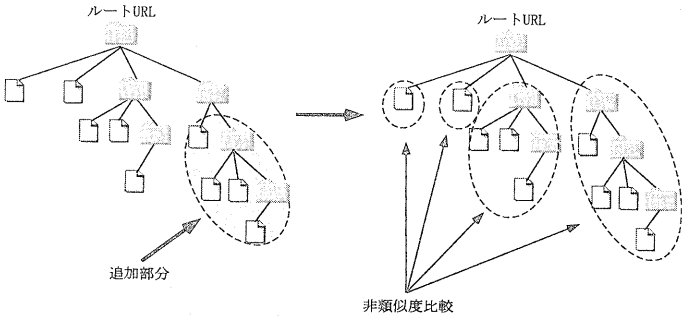

図 5　複数ページの内容比較

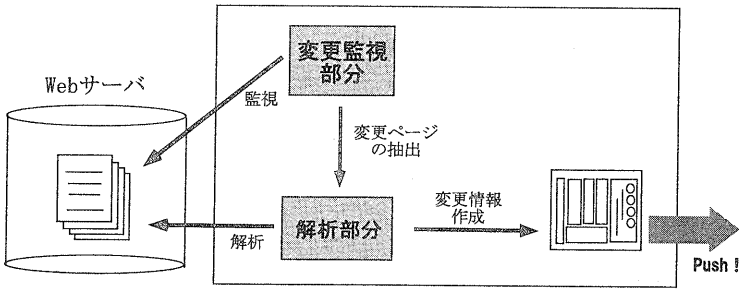

図 6　システム概要図

する．

このような情報呈示方式の 1 つとして，雑誌，週刊誌等の見出し広告（中吊り広告）の表現方式が考えられる．見出し広告にはその中に書かれてある情報の見せ方に差がある．例えばスクープ記事や特集記事などのいわゆる読む人に対してもっともアピールしたい情報ほど，最初に，目立つように見る人の目を引くようにレイアウトされている．見出し広告を見れば，そこにどんな情報があるか，どの情報をアピールしているかが一目で把握できるのである．

ページのレイアウトはあらかじめ設定しておき，各ページの評価順にそのページの情報を配置する．またここでユーザに対して実際に呈示する情報としては，そのページの内容を簡潔に表現し，またその概要を把握できるような情報が望ましい．このような情報として，各ページのタイトル，あるいは実際にページ内で見出しとして用いられているテキストを利用する．

次に，実際のユーザへの呈示方法であるが，当然のことながら作成した変更呈示のページを他のページと同様に Web ページとして公開するだけでは意味が無い．効果的にユーザへの通知を行うために，放送型情報配信を用いる．放送型の通知のメリットとしては，その

- 確実性
- 即時性

などが挙げられる．作成した変更呈示ページをプッシュ技術を用いてユーザに配信することで変更通知を行う．

5. プロトタイプシステム:Pudding

本章では今回試作したプロトタイプシステムについて述べる．このプロトタイプシステムでは仮想の Web サイトを構築し，そこで変更の監視・解析を行った．Web サーバとしては Apache1.3.12 を用い，Perl によりページの変更監視・解析部分を実装した．また，作成した変更呈示ページの配信には，プッシュ型配信システムである Castanet2.1[7] を用いている．

図 6 はシステムの概要図である．

変更監視を行う部分では，Web サーバ内に存在する全ページのタイムスタンプとファイルサイズの変更を

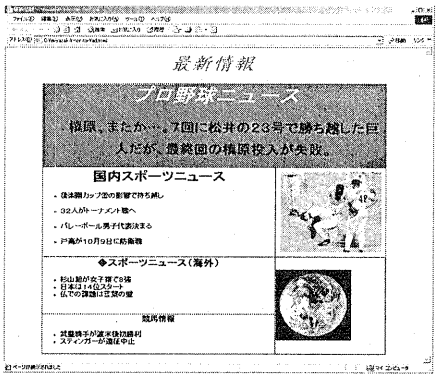

図 7　変更情報のサンプル画面 1

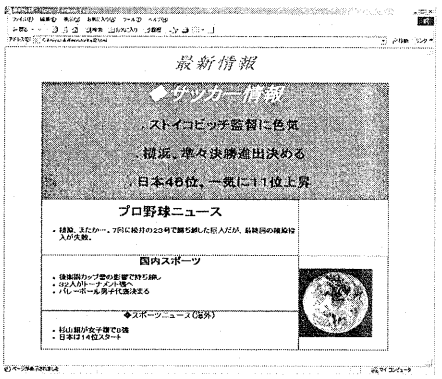

図 8　変更情報のサンプル画面 2

監視する．これらに変更が生じたもの，あるいは新しく作成されたページを検知した場合，解析部分がそのページに対して解析を行い，変更のランキングを再決定する．次に決定された各ページの価値に基ずいて，変更情報が作成される．

図 7，図 8 は作成されたサンプル画面である．これはスポーツ情報に関する仮想の Web サイトに対して，変更監視，解析から得られた変更情報である．これは図 7 は，このサイトに対するある時点の変更情報であり，ここからサッカーに関するページを追加すると図 8 の変更情報が作成された例である．追加されたサッカーに関する情報のページの評価値が，既存のページよりも高く評価され，その情報が変更情報の一番初めに表示されている．この変更情報に記述される内容は，各ページ内で実際に使われている見出しを表示した．

追加されたページは，

- 既存ページにサッカーに関するページが少ない．
- 新規ページの追加のため，アクセス数，更新頻度に関する評価値が，既存ページ内の最高値と評価された．
- 他の既存ページよりも，量 (単語数) が多い．

などの要因から，その評価値が高いと考えられる．各ページの評価値は，ページの追加，更新の際に算出され，時間の経過とともに減少するものである．

また，これらの変更情報のページは作成されると，自動的にユーザに配信される．ユーザ側では，変更情報が配信されると，ブラウザ内に配信された最新の変更情報が表示される．これによって，ユーザは常にサイトの最新の変更情報を知ることが可能である．

6. おわりに

本論文では，Web サイトの変更監視と変更の解析，そしてそれを基にした放送型の変更通知についての提案を行った．本研究の特徴としては次のようなものである．

- Web サイト全体を対象にした変更監視
  一つのページではなく，Web サイトが対象である．
- 内容の非類似度，変更量，アクセス数，更新間隔による変更の価値判断
  変更に対して，その価値を評価する．
- 価値評価に基づく放送型通知
  変更価値に基づいた変更情報の作成とプッシュ技術を用いた変更情報の配信．

今後の課題として，複数サイトの監視が挙げられる．複数のサイトを監視することで，最新の話題，サイト間のリンク関係など違った価値判断基準が考えられる．

さらに，ある変更を検知するたびにユーザへ変更情報を通知するのではなく，たとえば一日一回，一週間に一回というように，サイトの更新状況に応じた変更情報配信が考えられる．

またユーザプロファイルとの組み合わせなどにより，ユーザごとに個別化した変更監視，解析，変更情報の呈示などが考えられる．

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) 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),pp141-157(1999)．
2) WebCQ, http://www.cc.gatech.edu/projects/disl/WebCQ/
3) K. Takeda and H. Nomiyama, :"Site Outlining," ACM Digital Libraries (DL'98), pp. 309-310 (1998).
4) 武田浩一，中村祐一，浦本直彦：XML がもたらす創造的ネットワーク一動的な情報源と分散エージェント一，人工知能学会誌 Vol.14，No.6，pp.35-43(1999)．
5) WWWC：http://www.nakka.com/wwwc/
6) 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（TOD7，to appear）（2000）．
7) ローラ リメイ，：Marimba オフィシャルガイド Castanet，株式会社プレンティホール，(1997)．

盛光印刷所