マルチメディア通信と分散処理 85−43
（1997. 11. 7）

放送型データのユーザ適応型分類・選択手法

ソムヌック サグアントラクーン† 寺田 努† 塚本 昌彦† 西尾 章治郎†
三浦 康史‡ 松浦 聰‡ 今中 武‡

†大阪大学大学院工学研究化情報システム工学専攻
‡松下電器産業株式会社研究本部中央研究所

近年，多数の放送衛星の打ち上げにより，これらの衛星を用いたデータ放送サービスが提供されるようになった．このサービスによって提供されるデータは広範囲の分野にわたり，その量も膨大であることに対して，データを受信するユーザは一般に特定の分野に関するデータのみに興味をもつ．そのため，放送データを蓄える場合，受信した全てのデータを格納するのは非効率的であると考えられる．本論文では情報フィルタリング機構を放送型データ受信システムに導入することを提案する．フィルタリング機構を導入することによって，メモリの効率的な利用および格納されたデータへのアクセス時間の短縮が可能になる．また，フィルタリングの手法として木構造を用いてフィルタリングを行う手法を用いることでフィルタリングの結果として選択されたデータをユーザごとに特化して分類することが可能になる．さらに，本論文ではこれらの手法を用いて構築した放送型データ受信システムの設計・実装について述べる．

User Customized Classification and Selection for Broadcasted Data

Somnuk SANGUANTRAKUL† Tsutomu TERADA† Masahiko TSUKAMOTO† Shojiro NISHIO†
Kouji MIURA‡ Satoshi MATSUURA‡ Takeshi IMANAKA‡

†Department of Information Systems Engineering, Graduate School of Engineering, Osaka University
‡Central Research Laboratories, Corporate Research Division, Matsushita Electric Industrial Co., Ltd.

Recently, many broadcast satellites have been launched to provide data broadcasting services for public users. Although the service provided can cover many kinds of data and a wide range of users, it is considered that, in general, users will be interested only in a particular kind of data. Consequently, storing all data received is considered to be inefficient and only wasting a large amount of memory. This fact has strongly motivated us to introduce the information filtering system into a broadcast data receiving system. The use of filtering system will increase the efficiency of memory usage and reduce the access time of broadcast data stored in the system. In this paper, we propose a highly customizable filtering method that uses tree structures to represent user preferences. the use of this filtering method results in accurately classified data that suits the user's need. Further, we also describe the design and implementation of our broadcast data receiving system that makes use of the filtering method proposed.

1 はじめに

近年，多数の放送衛星が打ち上げられたことにより，これらの衛星を利用した放送サービスが利用可能となった．このような状況下で，従来のテレビ放送という利用法に加え，電子的なデータを放送するという利用法が注目されるようになった[2]．一般に衛星放送の帯域幅が非常に広いことから，これを利用したデータ放送では大量のデータ

を放送することが可能であり，さまざまな分野に渡るデータが放送されている．しかし，これらのデータ放送を利用するユーザは特定の分野にのみ興味をもつのが普通であるため，放送されるデータを蓄えたい場合，ユーザ側に大量なメモリを用意して放送された全てのデータを格納することは経済的ではない．また，膨大なデータの中から興味をもつデータを選択したり，データを整理したりすることもユーザにとって労力のかかる作業である．

このような問題を解決するためには，次のような機構が必要である

- 放送されるデータの中からユーザが興味をもつものだけを選択する機構．
- 選択したデータを分類する機構
- データの分類のしかたをユーザに特化する機構

本論文では，これらを実現するために放送データを受信するシステムに情報フィルタリング機構を導入することを提案する．また，フィルタリングを行う際に木構造を用いるフィルタリング手法を提案する．木構造を用いてユーザの嗜好を表現することで，ユーザは木を辿るだけで簡単にデータをアクセスできる．また，木の構造をユーザごとに特化することでユーザがより関心をもつデータをより迅速にアクセスすることが可能である．

以下では，まず，2章でユーザに適応する木構造を用いたフィルタリング手法について説明する．次に，3章で木の再構成について，4章で本研究で設計・実装を行った放送型データ受信システムについて述べる．最後に5章で今後の課題と本論文のまとめを述べる．

2 フィルタリング手法

本論文ではこれらを実現するために放送データを受信するシステムに情報フィルタリング機構の導入を提案する．従来の情報フィルタリングの手法はこれまでにも多数提案されているが[1, 3, 4, 6], Loeb の手法 [3] や Baclace の手法 [1] ではデータの分類については考慮していない．また，Stevens の手法 [6] では，放送源側におけるデータの分類とユーザ側の分類の間のギャップを埋めることができるが，ユーザが自分で分類方法を定義しなければならない．一方，データの分類に関しては小澤らが提案する背景知識を用いる手法 [5] があるが，ユーザの嗜好を考慮していないため，得られた分類木は必ずしもユーザにとって最適ではない．

本論文で提案するフィルタリング手法は放送データに対するユーザの嗜好を木構造で表現する．本手法では情報源側に放送データを分類するための分類木があると仮定し，この分類木をもとにユーザごとに特化する分類木を構築する．本論文ではユーザ側の分類木をカスタムツリーと呼び，これに対して放送側でのデータの分類木をグローバルツリーと呼ぶ．カスタムツリーが要求される機能は次のように挙げられる．

- ユーザの嗜好を正確に反映する機能．
- ユーザの嗜好の変化を迅速に追従する機能．
- ユーザの嗜好に合うデータを迅速に選択する機能．
- カスタムツリーの構成をユーザに特化する機能．

2.1 プロファイル適応度

まず，ユーザの嗜好の度合を表すために，プロファイル適応度を用いる．プロファイル適応度はユーザの嗜好を表す0から1までの実数値で，ユーザのアクセスパターンに基づいて自動的に計算する．プロファイル適応度には次の2種類がある．

節点のプロファイル適応度: カスタムツリーの各節点に付加されるプロファイル適応度であり，その節点が表す実世界の概念に対してユーザがどれだけ興味をもつかを表す．節点のプロファイル適応度はその節点に属しているデータの評価値から求められる．

データのプロファイル適応度: 放送されるデータを選択する際の基準としてデータのプロファイル適応度を用いる．あるデータのプロファイル適応度はそのデータに対してユーザがどれだけ興味をもつかを予想する値であり，データを受信した時点で計算される．

節点のプロファイル適応度を求めるためにデータの評価値を必要とする．データの評価値は，あるデータに対してユーザがどれだけ興味を示したかを表す値であり，ユーザの行動に基づいて算出される0から1までの実数値である．本手法ではデータの評価値は次式によって求める．

$$p = p_0 + p(t) \tag{1}$$

ここでpはデータの評価値であり，p_0はユーザがデータをアクセスしたことによって得た評価値であり，$p(t)$はアクセス時間による評価値である．ユーザが頻繁にデータをアクセスする場合，あるデータがアクセスされたことの特別さが低く，アクセスされたことによる評価値p_0も低いと考える．逆にユーザがまれにデータをアクセスする場合，あるデータがアクセスされたことの特別さが高く，p_0の値も高いと考える．ここではp_0は次式によって求める．

$$p_0 = (1 - r)w \tag{2}$$

ただし，rはデータの属している節点におけるアクセス率であり，wはその節点のプロファイル適応度である．

また，あるデータに対するユーザの興味はそのデータのアクセス時間に反映されると考える．ユーザがそのデータに対して興味をもてばデータのアクセス時間が長く，興味がなければデータのアクセス時間が低い．ただし，アクセス時間が非常に低い場合はユーザのミスアクセスと考え，アクセス時間があまりにも長い場合はユーザは途中でほかの作業をする可能性があり，アクセス時間による評価値の信頼性が低いと考える．アクセス時間による評価値$p(t)$は図1に示す評価関数によって

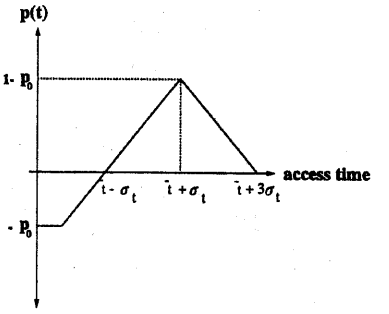

図 1: アクセス時間による評価値

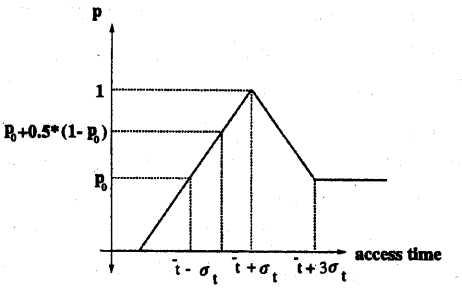

図 2: データの評価値

算出する．ただし，図1において$\bar{t}$はアクセス時間の平均値であり，σ_tはアクセス時間の標準偏差である．また，データの評価値pとアクセス時間の関係を図2に示す．

本手法では入力データに対するユーザの興味の度合はそのデータのプロファイル適応度で表す．入力データdのプロファイル適応度は次式によって求める．

$$\begin{aligned} P_d &= w_n + (1 - w_n)w_{n-1}P(c_n|c_{n-1}) + \cdots \\ &\quad + (1 - w_n)\ldots(1 - w_2)w_1P(c_n|c_1) \end{aligned} \tag{3}$$

ただし，$C_1, C_2, \ldots, C_{n-1}, C_n$はカスタムツリーの節点であり，節点$C_i$は節点$C_{i+1}$の親であり，$d$は節点$C_n$に分類されるとする．また，節点$C_i$の

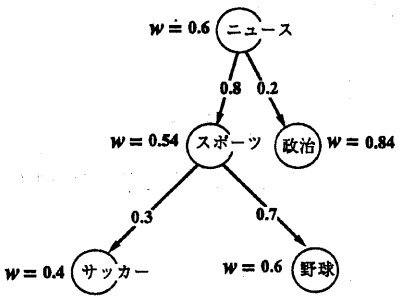

図 3: カスタムツリーの例

プロファイル適応度を w_i, C_iに分類されているデータの内, C_nにも分類されているデータの割合を $P(C_n|C_i)$ とする.

本手法ではデータのプロファイル適応度はユーザがそのデータをアクセスする確率であり, その値はデータのカスタムツリーにおける分類パス上の節点のプロファイル適応度に依存すると考える. また, 依存性は根に近い節点ほど小さくなると考える. (3) 式において, 右辺の第 1 項はデータが節点 C_nに属することによってアクセスされる確率であり, 第 2 項はデータが節点 C_nに属するがアクセスされず, 節点 C_{n-1}に属することによってアクセスされる確率である. このとき, $P(C_n|C_{n-1})$ は節点 C_{n-1}への依存性の減少を表す.

図 3にカスタムツリーの例を示す. ただし, 節点 C_iから節点 C_{i+1}への矢印は節点間の包含関係を表し, 矢印に付加される数値は $P(C_{i+1}|C_i)$ を表す. ここで, 節点「野球」に分類されるデータのプロファイル適応度は次式によって求める.

$$\begin{aligned} p &= 0.6 + (1-0.6) \times 0.54 \times 0.7 + \\ &\quad (1-0.6)(1-0.54) \times 0.6 \times (0.8 \times 0.7) \\ &= 0.6 + 0.15 + 0.06 = 0.81 \end{aligned} \quad (4)$$

3 木の再構成

本手法では一般に時間が経つにつれ, ユーザの嗜好の変化やカスタムツリーの状態の変化などにより以下のような現象が生じる.

カスタムツリーの誤差の増加: 一般にユーザの嗜好は外界などからの影響やユーザの状態の変化によって時間とともに変化するため, 現在ユーザが興味をもっているジャンルでも将来は興味をもたなくなることもあり, 逆に現在では興味をもたないジャンルでも将来は興味をもつ可能性がある. また, 同じジャンルでも時間が経つにつれ興味の度合いが変わってくることも考えられる. よって, 現時点でユーザの嗜好を正確に表すカスタムツリーでも将来はそうでなくなる可能性がある.

平均アクセス時間の増加: 時間が経つにつれ, プロファイル適応度の高い節点に格納されるデータ数が増加したり, 逆にプロファイル適応度の低い節点に格納されるデータ数が減少することがある. このようなカスタムツリーの偏りによって, ユーザの選択判定時間が増加したり, データへたどり着くまでのパスが冗長であることなどが生じ, データのアクセス時間が増加する可能性がある.

そこで, カスタムツリーを常に最適に保つためにはカスタムツリーの再構築を行う必要がある. 再構築には次の 2 つの処理を考える.

省略 (merge) 操作: 分類パス上の節点を省くことでカスタムツリーの高さを減らし葉節点までのパスを短縮する. 図 4に節点「野球」に適用する省略操作の例を示す.

細分化 (split) 操作: カスタムツリーの葉節点を細分化することで 1 節点あたりのデータ数を減らす. 図 5に節点「野球」に適用する細分化操作の例を示す.

また, 再構築時の制約条件として, データのプロファイル適応度の不変性を用いる. プロファイ

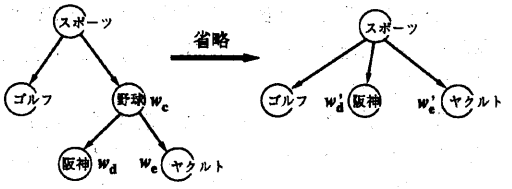

図 4: 省略操作

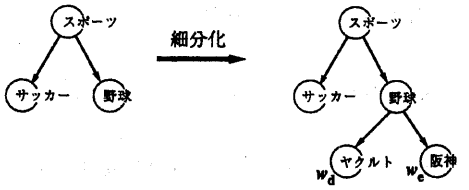

図 5: 細分化操作

ル適応度の不変性とは，再構築前と再構築後では，同じ節点に分類されるデータのプロファイル適応度は変わらないことである．この制約条件により図4では再構築後の節点「ヤクルト」および「阪神」のプロファイル適応度は次式によって求める．

$$\begin{aligned} w'_d &= w_d + (1 - w_d) w_c \mathrm{P}(d|c) \\ w'_e &= w_e + (1 - w_e) w_c \mathrm{P}(e|c) \end{aligned} \quad (5)$$

また，図5において，再構築後の節点「ヤクルト」および「阪神」のプロファイル適応度は次式によって求める．

$$\begin{aligned} w_d &= \frac{1 - \mathrm{P}(d|c) w_c}{1 - w_c \mathrm{P}(d|c)} \\ w_e &= \frac{1 - \mathrm{P}(e|c) w_c}{1 - w_c \mathrm{P}(e|c)} \end{aligned} \quad (6)$$

4 プロトタイプシステムの設計と実装

筆者らは衛星放送を利用した放送サービスに適用するシステムとしてアクティブ情報ストア (Active Information Store, AIS) の設計と実装を行っている．図6に本システムの概念図を示す．

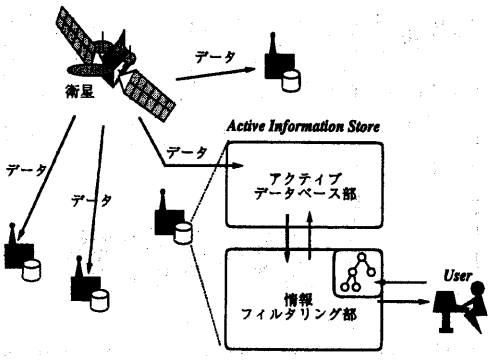

図 6: AIS の概念図

本システムで放送されるデータには放送側のグローバルツリーに基づいてデータの特徴を表す属性を付加される．ユーザ側ではこれらの属性に基づいてデータを選択する．ユーザ側のシステムは次のように分けられる．

情報フィルタリング部:
放送データに付加される属性に基づいてユーザの嗜好に合うデータだけを選択する．データの選択は2章に述べたような手法を用いる．

アクティブデータベース部:
選択されたデータを格納するためにスーパーアクティブデータベースシステム (SADB)[7] を用いる．SADB は放送サービスに適用するために拡張されたアクティーブデータベースであり，以下のような機能を備えている．

- 各データベース間でデータやパケットのやり取りができる．
- ECA ルールの送受信ができる．
- ECA ルールのグループ化や実行停止ができる．
- タイマ処理が可能である．

本システムでは衛星放送の代わりにネットワー

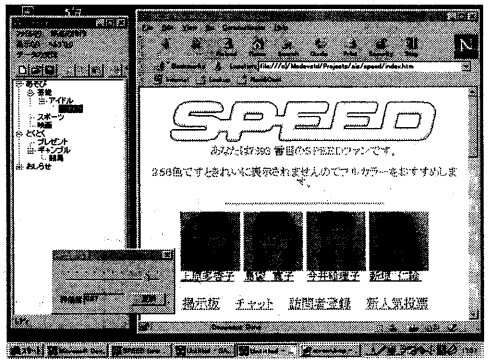

図 7: AIS の動作例

クを通してデータを放送するサーバを設ける．ユーザ側のシステムは Windows95 搭載のノートパソコン上で，Visual C++4.0 を用いて実装した．また，ユーザインタフェースとして，Netscape 社の Communicator4.03 ブラウザを用いた．図 7にシステムの動作例を示す．

本システムでは，SADB によって受信されたデータのカテゴリやサイズなどそのデータに関する情報に基づいてプロファイル適応度を求める．プロファイル適応度がしきい値以上のものだけを選択して SADB に格納要求を出す．また，ユーザが左の画面に表示されるカスタムツリーのデータを選択すると，フィルタリング部が SADB にデータ識別子を送り，返されたデータの詳細内容を右の画面に表示する．

また，時間の経過によって価値が少なくなったデータや期限切れのデータに対して，SADB は ECA ルールを用いて検出し，削除を行う．この場合，フィルタリング部に通知し，フィルタリング部はそれを受けてカスタムツリーを更新する．

5 おわりに

本論文では，放送型データ受信システムに適用する木構造を用いるデータの選択・分類手法について述べた．また，本論文で述べた手法を用いて構築した放送型データ受信システムの設計と実装について述べた．

木構造を用いてユーザの嗜好を表すことによって，ユーザが興味をもつデータを選択でき，選択したデータをユーザごとに特化するように分類できるようになった．また，ユーザは木を辿るだけで簡単にデータにアクセスすることが実現できた．

今後はシミュレーションを行って提案した手法の有効性の評価を行う予定である．

参考文献

[1] Baclace, P.E.: "Competitive agents for information filtering," *Comm. ACM*, vol. 35, no. 12, p.50 (Dec.,1992).

[2] 原島 博: "多チャンネル時代のコンテンツ製作," 日刊工業新聞社 (1997).

[3] Loeb, S.: "Architecting personalized delivery of multimedia information," *Comm. ACM*, vol. 35, no. 12, pp.39–48 (Dec.,1992).

[4] 森田昌宏: "情報フィルタリングに関する研究動向," JAIST Research Report, IS-RR-93-9I, 北陸先端科学技術大学院大学情報科学研究科 (Jun.,1993).

[5] 小澤順, 山田耕一: "背景知識を用いた概念学習によるデータベースからの知識発見," 人工知能学会誌, vol. 10, no. 6, pp.921–932 (Oct.,1995).

[6] Stevens, C.: "Automating the creation of information filters," *Comm. ACM*, vol. 35, no. 12, p.48(Dec.,1992).

[7] 寺田努, Sanguantrakul, S., 塚本昌彦, 西尾章治郎, 三浦康史, 松浦聰, 今中武: "アクティブデータベースを用いた放送型データ格納方式," 情報処理学会研究会報告 DPS(Nov.,1997, 発表予定).