

ユーザプリファレンス情報の自動生成と成長方式

Automatic Building and Tuning of User Preference Information for Fine Content Selection

喜田 洋二
Youji Kita亀山 渉†
Wataru Kameyama

1. はじめに

膨大なコンテンツの中からユーザの欲しい情報を得る手段として、個々のユーザの嗜好に合ったものを自動的に選択できることが望ましい。しかし、従来のアンケートなどの方法では、ユーザに労力をかけてしまうという問題がある。本報告ではユーザの嗜好、すなわち、ユーザプリファレンス情報（以下 UPI）を自動的に生成し成長させることでこの問題の解決する手法を提案する。具体的には、UPI をあらゆるパラメータをユーザの履歴からベクトルの形で抽出、その成長にベクトルの要素間の相関関係を反映させ、また、履歴がない状態に対しては、ユーザの属性ごとのモデルを生成する方式を検討しているのでそれについて報告する。

2. 提案方式概要

本研究ではメタデータを利用し、ユーザプリファレンス情報があるパラメータで表現することで、自動生成、成長が可能なものとしている。メタデータは TV-Anytime-Forum [1] で規定されているものを利用した。

2.1. 用語の定義

(1) Program Information Meta-Data (以下 PIMD)

コンテンツに付加されているメタデータ。そのコンテンツのタイトル、ジャンル、登場人物などの情報。本研究ではこれにその PIMD 自身の特徴をあらゆるパラメータを付加する。

(2) User Preference Meta-Data (以下 UPMD)

ユーザの属性情報。性別、年齢、職業などの情報。これは文献 [2] を参考に定義した。本研究ではこれに UPI をあらゆるパラメータを付加する。

(3) パラメータ

パラメータは3つ定義する。UPI をあらゆるユーザが保持する“UPI パラメータ”、PIMD に付加するその PIMD 自身の特徴をあらゆる“MD パラメータ”、他のユーザの UPI を参考にしたい場合に利用する“モデルパラメータ”である。ただし、これらのパラメータの構造は全て同一である。（詳細は § 3 を参照）

(4) フィルタリング

ここでいうフィルタリングとは、提案した UPI を用いてキーワード検索の結果をパーソナライズすることである。MD パラメータと UPI パラメータの類似度を用いて検索結果をソートして表示するものとする。

2.2. システム概要

実験環境としては図1のようなシステムで提案手法の検討を行っている。クライアントはサーバにコンテンツの検索要求を出し、サーバは要求にあった PIMD を返す、

というモデルを基本とし、以下のような条件の下に設計した。

1. UPI の生成、成長が可能であること
2. PIMD のフィルタリングが可能であること
3. モデルパラメータが生成できること
4. 動的に変化する PIMD の管理が可能であること

特に4番目の条件は、実サービスにおいてコンテンツが増え続けることを考えると必要不可欠な条件である。

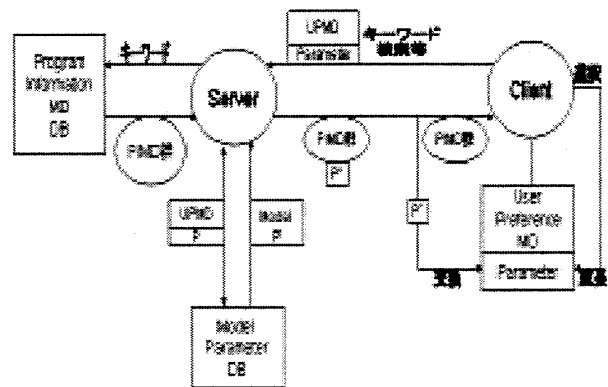


図1：システム概要

(1) サーバにおける定期的処理

サーバは定期的以下のような処理を行う。

- ・ PIMD を管理し、パラメータの構造を決定する。
- ・ PIMD をパラメータに変換し、それを PIMD に付加する。(MD パラメータ)
- ・ モデルパラメータを生成する。

(2) 全体の処理の流れ

全体の処理の流れはモデルパラメータの使用の有無により2通りある。

①モデルパラメータなし

まずクライアントがコンテンツの検索要求と UPMD (UPI パラメータ付加済み) をサーバに渡すと、サーバは

- ・ UPI パラメータの構造を更新する。
- ・ UPMD を Model Parameter DataBase に格納する。
- ・ 検索要求に合った PIMD を PIMD DataBase から引き出す。
- ・ 得た PIMD 群を UPI パラメータを用いフィルタリングする。

† Graduate School of Global Information and Telecommunication Studies, Waseda University

- ・更新した UPI パラメータとフィルタリングした PIMD 群をクライアントに返す。

という処理を行う。それを受けたクライアントは

- ・更新された UPI パラメータと現在保持している UPI パラメータを交換する。

という処理をし、ユーザが PIMD を選択すると

- ・選択した PIMD を選択履歴として蓄積し、UPI パラメータを成長する。

という流れになる。

②モデルパラメータあり

①と異なる部分としては、

- ・クライアントは検索要求と共にモデルパラメータの利用要求を出す。
- ・サーバは UPMD に合ったモデルパラメータを用い、PIMD 群をフィルタリングする。
- ・サーバはフィルタリングした PIMD 群のみを返す。である。

処理の詳細を次項から説明する。

3. パラメータ処理方式

3.1. 表現方法

パラメータはベクトルで表現する。ベクトルの各要素はコンテンツのタイトル、ジャンル、登場人物などの重要度をあらわす 0 から 1 までの数値とする。また、タイトル、登場人物をあらわすため次元数は冗長とする。

$$\text{パラメータ } \mathbf{p} = \begin{bmatrix} \text{TitleA} \\ \text{TitleB} \\ \text{GenreA} \\ \text{GenreB} \\ \vdots \end{bmatrix} \quad \text{ex,} \quad \mathbf{p} = \begin{bmatrix} \text{MATRIX} \\ \text{SEVEN} \\ \text{Action} \\ \text{SF} \\ \text{Comedy} \\ \text{Suspence} \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

3.2. 構造の決定方法

パラメータの構造はサーバが全 PIMD を参照し決定する。同時に各 PIMD の MD パラメータを生成し付加する。パラメータの構造は一意性を保つ。

3.3. UPI パラメータ生成、成長方法

UPI パラメータの生成は、初回検索時に得た PIMD の MD パラメータをそのまま UPI パラメータとすることで行う。

成長はユーザの検索履歴の蓄積と共に行われる。UPI パラメータに、選択した PIMD の MD パラメータを加算、平均し、偏りを得るため分散をとり UPI パラメータを更新する。さらにより新しい履歴に優位性をつけることで、履歴の蓄積による特徴の平坦化の問題に対応する。

3.4. モデルパラメータ

「ユーザはユーザ属性の類似した他のユーザが選択したコンテンツを選択する可能性が高い」という仮定のもとに、全ユーザの UPMD を属性でクラスタリングし、クラスタ内の代表パラメータをモデルパラメータとする。

クラスタリングはユーザ属性をベクトル化して行い、ユーザは自ユーザ属性と各クラスタ類似度を用いてモデルパラメータを得ることができる。

4. 相関関係の導出

パラメータベクトルの各要素は必ずしも直行関係にあるとは限らない。すなわち各要素間には相関関係があると考えられる。しかも、その相関値はユーザによって全く違う

ということも予想される。したがって、本研究ではユーザごとの相関値を算出し、パラメータに反映することでよりパーソナライズされた UPMD の生成を試みる。

相関値の算出は以下のような方法を提案する。

MD パラメータベクトルを \mathbf{md} とすると、ユーザの検索履歴は行列 \mathbf{PIMD} であらわすことができ、

$$\mathbf{PIMD} = (\mathbf{md}_1, \mathbf{md}_2, \dots, \mathbf{md}_n) = \begin{bmatrix} \mathbf{md}_{11} & \dots & \dots \\ \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \dots & \mathbf{md}_{nm} \end{bmatrix}$$

となる。この行列 \mathbf{PIMD} の行を要素にもつベクトル

$$\mathbf{md}'_n = (\mathbf{md}_{n1}, \mathbf{md}_{n2}, \dots, \mathbf{md}_{nm})$$

のコサイン相関値

$$\text{sim}(\mathbf{md}'_n, \mathbf{md}'_l) = \frac{(\mathbf{md}'_n, \mathbf{md}'_l)}{\|\mathbf{md}'_n\| \|\mathbf{md}'_l\|} = \cos \theta \quad \begin{matrix} l \neq n \\ 1 \leq l \leq n \end{matrix}$$

をとることにより、 \mathbf{PIMD} を履歴にもつユーザのパラメータ \mathbf{p} の、要素間の相関値行列 \mathbf{SIM} が得られる。

$$\mathbf{SIM}(\mathbf{md}'_n, \mathbf{md}'_l) = \begin{bmatrix} \text{sim}(\mathbf{md}'_1, \mathbf{md}'_2) & \dots & \dots \\ \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \dots & \text{sim}(\mathbf{md}'_n, \mathbf{md}'_{n-1}) \end{bmatrix}$$

\mathbf{SIM} はユーザの履歴 \mathbf{PIMD} に反映することで、そこから算出されるパラメータに影響する。また、フィルタリング時に利用することも検討中である。

5. まとめと今後の課題

本報告ではコンテンツ検索のパーソナライズに有用な、ユーザの嗜好、すなわち、ユーザプリファレンス情報を自動的に生成し成長する手法を提案した。特徴の 1 つは、ユーザプリファレンス情報をパラメータで表現することにより、嗜好の偏りを学習的に得てゆくという点である。

また、ベクトルの要素間の相関関係を得ることにより、より嗜好をユーザごとに個別化する方法を提案した。このような手法は空間の性質が未知な対象物に対する解析手法としても一般化できる可能性が考えられる。

今後本手法の有効性について検証するにあたり課題と考えられるのは、ベクトルの次元数とサンプル数の問題、相関関係算出方法、類似度算出方法であり、それぞれについて実験、検証を行う予定である。

【参考文献】

- [1] TV-Anytime Forum, "SP003v1.2", <http://www.tv-anytime.org/>
- [2] ISO/IEC JTC1/SC29/WG11, "MPEG-21 Requirements on Digital Item Adaptation", MPEG N4684