1T-1

# ウェブアクセスログを利用したユーザの嗜好抽出に関する考察

湯原　基貴†　　吉井　伸一郎†　古川　正志†

†北海道大学大学院　情報科学研究科

複合情報学専攻 複雑系工学講座自律系工学研究室

## 1. はじめに

ユーザのアクセス履歴を記録したアクセスログは有益なユーザ情報を取得する手段として注目されている。これまでにアクセスログを利用してユーザの行動パターンや傾向を抽出する研究が行われてきた[1]。しかしアクセスログはウェブサーバに設置されたサイトに対するアクセスのみ記録する。また動的 IP アドレスでアクセスするユーザを一意に特定することはできない。このため WWW 上のユーザ行動を分析することは困難である。

本稿では、ユーザの嗜好を抽出する対象として、無作為に抽出されたユーザを対象にアクセス履歴を収集したインターネット視聴データを利用する。インターネット視聴データは一意のユーザ ID に基づいてユーザの PC からアクセス履歴を取得するためサイト間の遷移も記録されている。

ユーザがアクセスした全てのウェブページに興味をもっていると仮定し、ベクトル空間モデルを用いて、そのウェブページが属するドメインに対するアクセスを要素としたベクトルによってユーザの嗜好を解析する。

## 2. インターネット視聴データ

インターネット視聴データは、数字を無作為に組み合わせて調査対象の電話番号を作成する RDD 方式による調査で PC ユーザ数を推定し、その結果を基に地域毎のモニター数を決定している。このため統計的に代表性をもったデータである。

### 2.1. インターネット視聴データの構成

インターネット視聴データには主要な項目として、

- ユーザ ID
- ウェブページにアクセスした時刻
- ウェブページを参照した時間
- ウェブページのドメイン名
- ウェブページの URL

がある。

本稿で用いたインターネット視聴データの概要を表１に示す。

**表 1　インターネット視聴データの概要**

| | |
|---|---|
| **期間** | **06/11/6〜06/11/12** |
| **ユーザ数 (人)** | **5519** |
| **総アクセス数 (件)** | **3702919** |
| **アクセスした URL** | **1581867** |
| **アクセスしたドメイン** | **48898** |
| **平均参照時間** | **25.8** |

## 3. インターネット視聴データからの嗜好抽出

ユーザは興味のある情報を選別してアクセスを繰り返している。参照時間とウェブページの評価値の相関は高いことが報告されているが[2]、個々のユーザで変動が大きいと考えられる。そのため本稿ではアクセス回数をユーザの URL に対する興味の大きさとする。

### 3.1. ドメインを成分とする嗜好ベクトル

URL に対するアクセス数を要素とするベクトルをユーザの嗜好を表現した嗜好ベクトルと考える。このような高次元ベクトル空間による表現は情報検索の手法の一つとして知られている。

しかし、URL に対するアクセスを要素にした場合、表 2 より約１５８万という非常に高次元なベクトルとなり計算が困難である。そこで、ドメインを、互いに関係性のある類似した内容をもつウェブページ（URL）の集合であると考え、ドメインに対するアクセスを要素とする嗜好ベクトルを構築する。次元数は約５万次元であり、同じドメインに属する URL に対するアクセスは同じ次元で表現される。ユーザ $i$ の嗜好ベクトル $V_i$ を次の式に表す。

$$Vi = \left(n(d_1), n(d_2), ..., n(d_{|D|})\right) \quad (1)$$

$n(d_i)$ はドメイン $d_i$ に対するアクセス数、$D$ はドメインの集合を現す。また、アクセス数の違いでベクトルの大きさが変化するのを防ぐため以下のように正規化を行う。

$$\|V_i\| = \frac{Vi}{|Vi|} \quad (2)$$

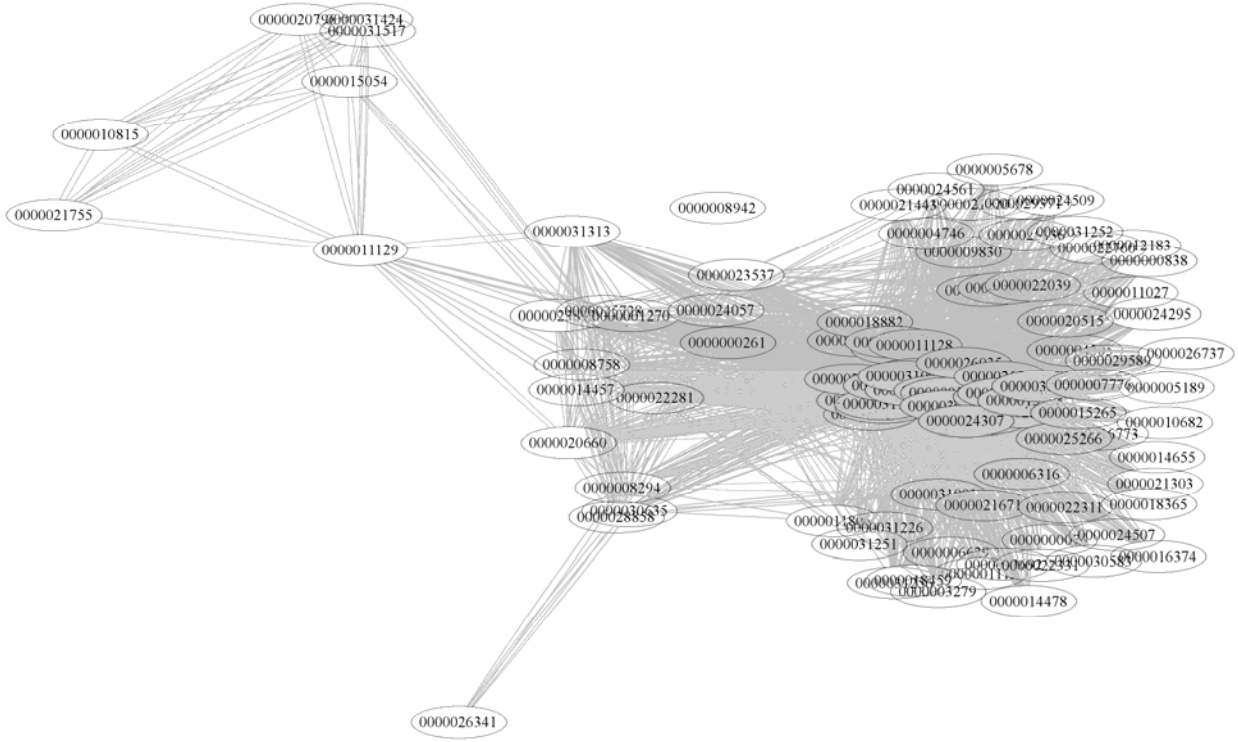


**図 1 ユーザの関係を表したグラフ**

## 4. 嗜好の類似度の計算

ユーザの嗜好をベクトルで表現したことで、ベクトル空間上で定義される距離によってユーザ間の嗜好の類似度を計算することができる。

### 4.1. コサイン係数

コサイン係数により定義される距離を利用してユーザ間の類似度を決定する。

ユーザ $i$ とユーザ $j$ と類似度 $sim(i, j)$ は嗜好ベクトル $V_i, V_j$ を用いて次の式で表される。

$$sim(i, j) = \cos(V_i, V_j) = \frac{V_i^T V_j}{\|V_i\| \cdot \|V_j\|} \quad (3)$$

### 4.2. 類似度によるユーザの関係のグラフ化

アクセス数上位 100 ユーザ間の類似度を計算し、ユーザをノード、$sim(i, j) \geq 0.5$ であるユーザ間をリンクで結び、ユーザの関係をグラフ化した（図１）。グラフは類似度が高いノード間ほど距離が近い。図１より類似度が高く密接したユーザ集合がいくつか存在することが分かる。属する集合内のユーザと嗜好が近く、それ以外のユーザとの嗜好が異なると考えるとユーザの嗜好が相対的に抽出できる。

## 5. まとめ

ユーザの興味がアクセスした全ての URL にあると仮定して、ユーザの WWW 上でアクセスを記録したインターネット視聴データから関連性の高い URL 集合に対するアクセスを要素とするベクトルを作成し、ユーザの嗜好を表現した。また、ベクトル間の距離によってユーザ間の嗜好の類似度を計算し、ユーザの関係を表すグラフを作成することで嗜好の類似したユーザの集合を発見した。本稿ではベクトルの成分である URL 集合としてドメインを利用したが、今後はクラスタリング手法などを用いて、より関連性のある URL 集合を作成する予定である。

## 6. 謝辞

本研究で利用したインターネット視聴データは、（株）ビデオリサーチインタラクティブ社五十嵐達様、江村健太郎様にご提供戴きました。ここに感謝の意を表します。

## 7. 参考文献

[1] Batisa,P.and Silva, M.J:Mining on-line newspaper web access logs,12th international Meeting of the Euro Working Group on Decision Support Systems,May 2001

[2] M. Claypool, P. Le, M. Waseda, and D. Brown. Implicit interest indicators. *Intelligent User Interfaces (IUI)*, pages 33–40, 2001.