社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－67 (16)
2007－BIO－11 (16)
2007／12／20

特定ユーザのための嗜好分析パターン抽出の一手法

児玉 理映子 石川 千里 高田 雅美 城 和貴
rieko425@ics.nara-wu.ac.jp
奈良女子大学 大学院人間文化研究科 情報科学専攻

概要

ユーザの嗜好に合致する商品を提示するレコメンド法が注目されている．しかしながら，従来のレコメンド法では全てのユーザを満足させることはできない．そこで，満足できないと考えられる特定ユーザの嗜好を抽出する新たなレコメンド法を提案する．本稿では，特定ユーザとして興味の有無が明確なオタクを採用した．オタクの嗜好を抽出するにあたってまずオタクの定義をし，その定義に基づいて選出したユーザをオタクとする．またオタク以外のユーザを一般人とする．このオタクと一般人のWebアクセスログから決定木を生成しネット上での行動パターンを抽出し分析する．分析結果からオタクは一般人と比べ膨大な情報量を持つコンテンツを閲覧していることが示された．

A Pattern Extraction Method to Analyze Favorite of Specific Users

Rieko Kodama, Chisato Ishikawa, Masami Takata, Kazuki Joe
Faculty of Sciene, Nara Women's University

Abstract

Recommendation methods that offer goods to users according to their favorite attract attention recently. However, existing recommendation methods cannot be applied to all users in even. So, we propose a new recommendation method that extracts specific users who are not happy with recommended goods. In this paper, we adopt Otakus, whose interest is clear, as specific users. First, for extracting Otaku's favorite, Otaku is defined, and several users selected according to the definition are regarded as otakus. The access pattern on the Internet is extracted from web access log of the otakus and other standard people to be analyzed. The result shows that otaku browses more copious contents than other standard people.

1 はじめに

近年，様々なオンラインストアが増加している．それに伴ってその利用者数も年々増している．

一部のオンラインストアでは，ユーザの嗜好に合致する商品を提示するレコメンド法などが採用されている．レコメンド法を取り入れた機能のことをレコメンデーション機能と呼ぶ．通販サイトであるAmazon.comではこのレコメンデーション機能の実用レベルの最先端を走っている [1]．例えば，トップページには「おすすめ商品」や「サーチ履歴から関連商品のおすすめ」,「あなたへのおすすめ」などを表示するようになっている．しかし，従来のレコメンド法は，一般ユーザの購買パターンから嗜好を抽出しているため，全てのユーザを満足させることはできないと考えられる．そこで，我々は従来のレコメンド法では満足できないと考えられる特定ユーザに着目し，そのユーザの嗜好を抽出する．

本研究では，特定ユーザとして興味の有無が明確であり，独特の購買パターンを持っていると考えられるオタクを採用する．野村総合研究所の調査によると，オタクは12分野に分類されており，2004年の市場規模は全分野で4110億円となっている [2]．よって，オタクはマーケティング戦略上無視できない存在となっており，オタクの購買パターンを解析したレコメンド法を開発すれば，さらなる市場拡大に繋がると考えられる．本稿では，まずオタクの定義をし，定義に該当する被験者をオタクとし，それ以外の被験者を一般人とする．被験者の分類を行った後，被験者のWeb上でのアクセスパターンを抽出する．さらに，抽出結果からオタクと一般人のWeb上での行動パターンを比較しオタクの行動を分析する．

以下2章ではパターン抽出方法について述べ，3章ではWebアクセスログの前処理，4章では決定木の生成，5章では本研究のまとめと今後の課題について述べる．

2 パターン抽出方法

2.1 オタクの定義

本研究では，特定ユーザとしてオタクを採用した．

オタクは情報収集が早いという特徴を持っており，従来のレコメンド法で提示された商品ではすでに購入または，既知である可能性が高く，ゆえに従来のレコメンド法で提示される情報だけでは購買促進効果を十分に得ることができない．よって，オタクの独特な購買パターンに一般的なデータマイニングによるパターン抽出は適応できず，正確なオタクの嗜好を解析することはできないと考えられる．オタクへの適応を可能にするためには，オタクの独特な行動パターンを分析する必要がある．この分析を行うためにはオタクを定義しなければならない．本稿では，オタクは強くこだわりをもっている分野が明確で，一般人以上にその分野に対する追求性が高く，趣味や余暇として使える時間のほとんどすべてをその分野に費やす特性をもっていると定義する．また，興味のない分野に対して，オタクは全く関心を持たない．オタクは興味のあることに全てを捧げるため，衣食住といった一般人にとっては欠かせない物事であっても興味がなければ対応する意思が全くないものだと定義する．

2.2 パターン抽出の流れ

まずパターン抽出を行う前に被験者を決定する必要がある．被験者には上記のオタクの定義に該当するオタクである人を選び，それ以外の被験者を一般人とする．本稿では，まず容易にデータを取得することができるWebアクセスログを用いて，ネット上での「オタク」と「一般人」のアクセスパターンの分析を行う [4]．この分析結果から，本稿ではWeb上でのオタクの行動パターンを抽出する [5]．分析手法としてデータマイニングを採用する．

2.3 パターン抽出の手法

本稿で提案する手法について説明する．

まず，各被験者のWebアクセスログデータを取得し，そのデータに対して前処理を行う．次に，前処理を行ったデータに対してデータマイニングを行い，Webサイトへのアクセスパターンの特徴を抽出する．今回事前にオタクであるか一般人であるかを2.1節にて述べた定義に基づき選出した被験者から「オタク」と「一般人」の特徴を抽出するのが目的である．そのための手法として，本研究ではデータマイニング分野において予測モデルであり，ある事項に対する観測結果からその事項の目標値に関する結論を導く「決定木」を採用する．

3 Webアクセスログの前処理

本研究では，Webアクセスログを取得するにあたってプロキシサーバsquidを利用する [6][7]．Squidのプロキシログはインターネット利用者のWeb上での行動履歴を時間順に保持したものである．今回研究で使用した主要なプロキシログデータは以下の通りである．

- Proxy data[Stay Time, IP, Result Code, Byte, URL, Type]

StayTimeはキャッシュの読み込み時間，IPはリクエストしてきたクライアントのIPアドレス，Result-CodeはキャッシュのIP要求の結果コード，Byteはクライアントへ渡された総バイト数，URLはリクエストされたURLが記憶されている．またTypeはHTTPヘッダーにリプライされてきたオブジェクトのコンテンツタイプを表している．本節では，このプロキシログデータに対して適用する前処理について述べる．

3.1 各被験者別のアクセス履歴作成

プロキシログは全ての被験者の行動履歴を保持している．このプロキシログを被験者IDごとにデータセットを分割することによって被験者別のアクセス履歴を作成することができる．被験者別のプロキシデータの前処理は図1の手順で行う．

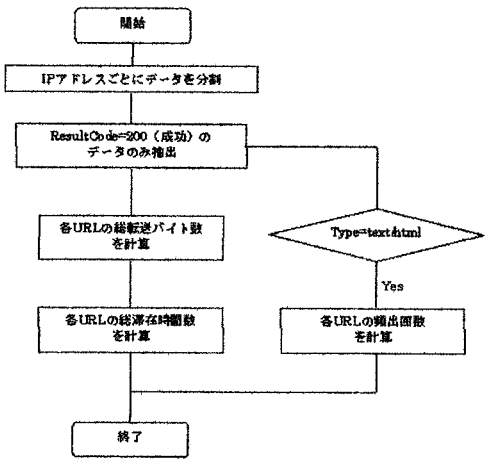

図 1: 前処理の手順

まず，各被験者のIPアドレスごとにデータセットを分割する．次に，実際に閲覧することができたＵＲＬを含むデータセットが必要となるので，分割したアクセス履歴に対してResultCodeが閲覧成功をしているデータセットを抽出する．抽出したデータは[StayTime, Byte, URL, Type]となっている．さらに，各被験者のアクセス閲覧から同一URLをまとめ，各URLに対して閲覧頻度数[count]，総転送バイト数[all_byte]，総滞在時間数[all_staytime]を算出する．各ＵＲＬの閲覧頻度数に関しては「Type=text, html, xml」の場合のＵＲＬデータ数を算出して求める．

総転送バイト数の算出方法に正規化後の各URLの転送バイト数と各URLの閲覧頻度数の積を利用する．転送バイト数の正規化については3.2節で説明する．

上記の過程によって生成されたるデータ[count, all_staytime, all_byte]となる．この生成した被験者別のデータに対してデータマイニングを行いWebサイトへのアクセスパターンを抽出する．

本稿で協力して頂いた被験者は，本研究室の学生14名である．被験者のプライバシーを尊重し個人は特定できないようにしている．

3.2 転送バイト数の正規化

総転送バイト数を算出するために各URLの転送バイトに対して正規化を行う．本稿では正規化にシグモイド関数を使用した．シグモイド関数は次式で表される．

$$\frac{1}{1+e^{ax}}$$

本稿では$a = 1 \times 10^{-4}$とし，またxは転送バイト数である．

4 決定木の生成

4.1 実験方法

本稿では，決定木を生成する上で，今回決定木アルゴリズムJ4.8を使用する[3]．被験者はオタク6名，一般人8名である．実験データ，および検査データに関しては以下の表1の通りである．実験回数が1回では最良な結果を得ることができないと考え，オタク6名から3名，一般人8名から4名を選出する全ての組み合わせに対して実験を実施する．実験パターンは1400となる．

表1の被験者の各データは3.1節で算出した通りである．決定木生成に使用したデータは以下の通りである．

[count, all_staytime, all_byte]

count, all_staytime, all_byteに関しては3.2節で述べたように，countは各URLの閲覧頻度数，all_staytimeは各URLの総滞在時間数，all_byteは各URLの総転送バイト数を表している．また，ここでのtypeは被験者の種類を表しており，「オタク」か「一般人」のどちらかが記録されている．また，各被験者のall_byte降順10位までのデータを使用した．実験1回の使用実験データ数はオタク30，一般人40で，使用検査データ数はオタク30，一般人40である．

被験者	オタク	一般人
実験データ	3	4
検査データ	3	4

表1: 被験者のデータ数

4.2 実験結果

実験データに対して決定木アルゴリズムJ4.8を適応した結果，正答率70 %以上の分岐条件を選出した．さらに，この分岐条件に対して検査データを適合した．その中でも正答率最大のものを採用する．採用された分岐条件は以下の通りである．

```
分岐条件
count <= 1: オタク (39/60)
count > 1
|  count <= 4
|  |  all_staytime <= 4953
|  |  |  count <= 3: 一般人 (19/80)
|  |  |  count > 3
|  |  |  |  all_staytime <= 1360: 一般人 (1/80)
|  |  |  |  all_staytime > 1360: オタク (2/60)
|  |  all_staytime > 4953: オタク (6/60)
|  count > 4: 一般人 (23/80)
正答率：64.2857143%
オタク：70% 一般人：57.5%
```

上記で示した分岐条件の()内は条件に適合しているデータ数を表しており，正答率は検査データ適合後の数値となっている．

上記の分岐条件より，オタクと一般人のデータに対して比較を行う．比較方法は以下の通りである．

(1) オタクと一般人それぞれのデータ数全体に対して適合率が最大の条件分岐を選出し枝狩りをする

(2) (1) から選出されたオタク・一般人の各データ (count, all_staytime, all_byte) の平均値を求める

(3) (2) で求めた平均値からさらに 1 回のアクセスに対する平均転送バイト数・平均滞在時間数を計算する

(4) (3) の平均値をオタク・一般人の各データに対してそれぞれ比較を行う

比較方法 1 より，オタクについての適合率最大の条件分岐 $count \leq 1$ を採用する．これはデータ数全体 (データ数：60) の約 51.5%(データ数：39) を占めている．同様に，一般人についての適合率最大である条件 $count > 4$ を採用した．これはデータ数全体 (データ数：80) の約 28%(データ数：23) を占めている．各データから求められた平均値の計算結果は表 2 の通りである．

1 アクセス	平均転送バイト数 (正規化)	平均滞在時間数 (s)
オタク	2.56E-02	1890(約 32 分)
一般人	9.03E-62	810(13.5 分)

表 2: 平均転送バイト数と平均滞在時間数

比較した結果，一般人の平均転送バイト数に対してオタクは 2.84E+59 倍の転送バイト数を持つホームページを閲覧している．平均滞在時間に関して，オタクは一つのコンテンツに対して一般人の 2.34 倍の時間を要して閲覧している．

4.3 考察

比較結果からオタクは一般人と比べ膨大な情報量を持つコンテンツを閲覧していることがわかる．またそれに伴い閲覧時間も一般人と比較して長く，平均閲覧時間は約 32 分となっている．これより，オタクが閲覧しているサイトは転送バイト数が大きく閲覧時間を長時間要する動画やブログなどのサイトを閲覧していることが考えられる．一方，一般人は転送バイト数が少ないサイトを幾度も閲覧していることから検索サイトなどの目的を持ったサイトを閲覧することが多いと推測される．

さらに，本稿で定義したオタクの被験者によって生成された決定木を利用してオタクの嗜好を抽出することで，オタクが必要としない情報を除くレコメンド法を適応することができると考えられる．

5 おわりに

本研究では，従来のレコメンド法では満足できないと考えられる特定ユーザのためのレコメンド法の必要性に着目した．

レコメンド法ではユーザの嗜好を分析する必要がある．本稿では，一般人とオタクの Web アクセスログを利用して Web 上でのアクセスパターンを比較し，特徴を抽出した．さらに，この特徴抽出からオタクの Web 上での行動を分析した．分析結果から，オタクは一般人と比べ膨大な情報量を持つコンテンツを閲覧しており，また閲覧時間も一般人より長いことから動画やブログなどのサイトへアクセスする機会が多いことが示された．また，本稿で生成した決定木を利用して分類されたオタクの嗜好を抽出することで，膨大な商品の中からオタクが求める商品を絞り込むレコメンド法が実現可能であると考えられる．

今後，生成した決定木が何万人というインターネットユーザに適応できるかを検証する必要がある．同時に，オタクに分類された被験者の嗜好を分析していく予定である．

参考文献

[1] 松本晃一:"アマゾンの秘密", ダイヤモンド社 (2005)

[2] 野村総合研究所 (NRI):NEWSREREASE(2004) http://www.nri.co.jp/news/2004/040824.html

[3] Weka the University of Waikato: http://www.cs.waikato.ac.nz/ ml/weka/index.html

[4] 山田和明，中小路久美代，上田完次:"ウェブ・アクセスログに基づくインターネットユーザの興味遷移パターンの抽出", ロボティクス・メカトロニクス講演会， Vol.2005(20050609) p. 178

[5] 山田和明，中小路久美代，上田完次:"Web ユーザの行動履歴解析のためのデータマイニング", 電子情報通信学会第 3 回 Web インテリジェンスとインタラクション研究会

[6] squid-cache.org: http://www.squid-cache.org/

[7] 成凱, 平野真太郎, 上林弥彦:"プロキシログ解析に基づくトップページの抽出と検索", 電子情報通信学会第 14 回データ工学ワークショップ (2003)