

ユーザの要求変化に着目したウェブ閲覧履歴の分類方式

長野 翔一^{†1} 高橋 寛幸^{†1} 中川 哲也^{†1}

情報爆発時代において、情報の個人化を実現するプロファイル技術が注目されている。しかし、現在のプロファイル技術は獲得した閲覧履歴全体からユーザの全閲覧行動における要求の傾向を推測するため、要求の変化を検出するのは困難である。本稿が扱う「要求」とは行動への動機の事を指し、10分程度で変化する性質を持つ。ユーザは要求に基づいてウェブページの閲覧を行う。

我々は、ユーザの要求変化は各閲覧履歴の類似度を利用することで検出可能であると考え、閲覧履歴の分類方式を提案する。既存の分類方式では、同じ要求内でも時系列に従い少しずつカテゴリが変化する、複数の異なる要求が並存する、といった閲覧行動の性質のため精度を下げることとなる。そこで、提案方式はこれらの性質を考慮し、クラスタ重心付近に十分な閲覧履歴数が確保できないことを前提とした、局所解重視の分類方式の構築に取り組む。

また、既存の分類方式と比較実験を行い、提案方式が既存方式に比べ有効であることを確認した。

Clustering method for detecting dynamic intentions from browsing behavior

SHOUICHI NAGANO,^{†1} HIROYUKI TAKAHASHI^{†1}
and TETSUYA NAKAGAWA ^{†1}

We propose a clustering method for detecting the change of intention from user's browsing behavior. It is necessary to treat the user's intention accurately in information explosion age. However, treating dynamic intention is difficult for a conventional method, as behavior targeting model. Because the category change little by little in the same intention, and any other intentions exist at same time.

For detecting user's intention change in browsing-behavior, we analyze each of browsing-history based on the similarities, and clustering based on local part similarities, in case web history have not a normal distribution.

In addition, we evaluate on result of an experiment to effectiveness for conventional clustering method.

1. はじめに

情報爆発時代において、ウェブ上では様々な情報の個人化技術 (amazon のレコメンデーション¹⁾ など) が創出されている。情報の個人化を実現するためには、性別や年代といったデモグラフィック情報や、ユーザがある時点でどのような情報を求めているかを把握するプロファイル技術が不可欠である。しかし、デモグラフィック情報から推測される情報は平均的なユーザグループの傾向を示すため、個々のユーザの興味を推定できないという問題があった。

近年、ウェブ上では行動ターゲティングをはじめとして、閲覧履歴 (閲覧番号、時間、タイトル、URL 等を時系列順に並べたデータ) を利用して、個々のユーザの興味を把握するプロファイル技術が注目されている。

「興味」とは1ヶ月程度持続する行動への動機を指し、山田 (2005)³⁾、戸田 (2007)⁴⁾ など、興味を対象とした研究は数多く行われている。一方、本稿で扱う「要求」とは、ユーザがある時点における行動への動機の事を指し、頻繁に出現する要求が興味である。これまでの実験²⁾ の結果から、要求は10分程度の期間持続する性質をもつことが分かっている。

既存のプロファイル技術は、ユーザの興味や要求が履歴上に反映されることを前提として、閲覧履歴を時系列順に獲得し、獲得した閲覧履歴全体から出現頻度の高いものをユーザの閲覧行動における興味として抽出する。そのため、頻繁に変化する性質を持つ要求を対象とする場合、ユーザの要求変化に対応することが困難である。

そこで、我々は、要求の変化を閲覧履歴の分析から検出することが可能であると考え、要求から生成された閲覧履歴を要求ごとに閲覧履歴を分類する要求分類方式を提案する。

提案する分類方式がプロファイル構築において、どのように利用されるかについて以下に示す。

図1はあるユーザの閲覧履歴である。閲覧履歴は左から、時系列順に並んだID、ウェブページタイトル、要求が変化した閲覧履歴、並行して行っていたセッションの識別子で構成されている。この閲覧履歴前半においては、奈良地域に関する要求を有しているが、ID.61の中央新幹線をきっかけに奈良地域に関する要求が電磁気学に関する要求に変化し、後半においては、電磁気学に関する要求と化学に関する要求が並存している。

この閲覧履歴に対して本稿で提案する分類方式を適用すると図1のように要求x、要求y、要求zの3つの要求から生成された閲覧履歴に分類される。そのため、例えば直近の要求を取り出す際に、奈良地域に関する要求を除去する、電磁気学に関する要求を分析する際、並存する化学に関する

^{†1} 日本電信電話株式会社 NTT 情報流通プラットフォーム研究所
NTT Information Sharing Platform Laboratories, NTT Corporation

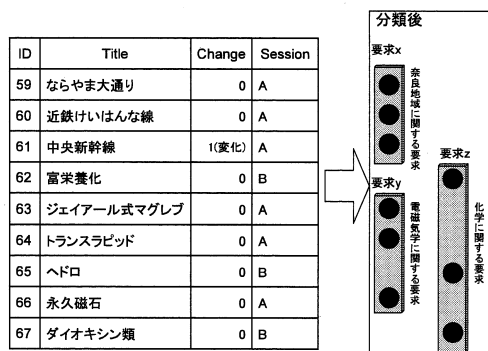


図 1 評価対象となる閲覧履歴の一例

る要求を除去するといった処理が可能となり、プロファイル構築の精度を向上させることができる。

本稿の構成について以下に説明する。

はじめに、2 章において背景、研究が取り組む課題について示す。3 章において提案方式を説明し、競合への優位性、研究の位置付けについて示す。4 章において提案方式の有効性を検証した評価実験について示す。最後に 5 章においてまとめについて示す。

2. 背 景

2.1 要求分類の必要性

プロファイル技術は長期にわたる興味の抽出を想定しており、頻繁に変化する要求の抽出を想定していない。そのため、要求変化が起こればプロファイルの精度は低下する。

これまででは、要求変化に対応するため、変化点などを検出し、期間ごとに閲覧履歴を分割することで要求の変化を抽出していた。しかし、複数の異なる要求が同時に存在したり、分割した期間内で要求変化が起これる場合、期間ごとの分割では、複数の要求から生成された閲覧履歴が一つの期間に混在し、要求変化に対応できない。

プロファイル技術が要求変化に対応するためには、1 ユーザの閲覧履歴から要求を推定するのではなく、閲覧履歴をユーザの要求ごとに分類し、分類された閲覧履歴からプロファイルを構築する必要がある。これにより、閲覧履歴上に表出しなくなった要求を閲覧履歴上から検出することが可能となるため、要求の変化に対応できる。また、ユーザが同時期に複数の要求を持ち、閲覧履歴上で混合し、配置されていた場合も、処理中の要求に無関係な閲覧履歴を除去することができる。

2.2 履歴分類によって変化を捉える既存方式の問題点

閲覧履歴の分類やクラスタの関係抽出に関する研究は数多く行われているが、10 分という短期間しか持続しない要求の変化を抽出するための閲覧履歴分類は、要素となる閲覧履歴数が少なくなるため、困難である。

既存方式には閲覧履歴を分類することで長期にわたる興味遷移を捉える方式がある。例えば、山田 (2005)³⁾ は x-means 法による分類を提案している。これは、ウェブページの特徴値 (単語と重要度をベクトルとし、単語ベクトル

を主成分分析にかけた主因子) を時系列にソートすると正規分布を有するという仮定に基づき x-means 法⁵⁾ を利用した分類を行ない、クラスタの特徴値の変化を利用して長期的な興味遷移を捉え、それを可視化する方式である。

しかし、要求変化を捉える閲覧履歴分類を行う場合、クラスタの要素数が減少し、x-means 法的前提である正規分布が成立しにくくなり、分割数設定の精度が下がるという問題が発生する。

そこで本稿では、要求の変化を明らかにするために、要求の性質が閲覧履歴上に表出すると考え、ユーザの要求の性質が閲覧履歴に与える影響を考慮した分類方式の構築に取り組む。

3. 閲覧履歴の分類方式の提案

3.1 提案内容

本章では、閲覧履歴から取得したウェブページ本文の類似性を利用し、生成した要求ごとに閲覧履歴を分類する要求分類方式を提案する。

我々は、短期的な要求が次の 2 つの性質を有するため、閲覧履歴上の特徴として反映され、分析が困難になると考える。そのため、これらの性質を考慮した要求分類方式を構築する。

研究課題 1 同じ要求が生成した閲覧履歴でも、時系列に従い少しずつ要求が変化している

研究課題 2 複数の異なる要求が並存することがある。

そこで、本稿では上記 2 つの性質を利用し、「局所解重視のクラスタリング」と「類似度による既成クラスタへの要素組み込み」を順に行う 2 段階の要求分類方式をアルゴリズムに組み入れる。

3.2 アルゴリズム

本項では、提案方式のアルゴリズムを紹介する。処理は以下の処理 1、処理 2 を経て分類結果が出力される。処理 1 で確実に同じ要求から生成されたものをまとめてクラスタの基礎を作り、処理 2 では処理 1 でクラスタの要素とならなかった閲覧履歴をクラスタに振り分ける処理を行っている。

入力となるのは各閲覧履歴に対応した本文の類似度を記述したマトリクス表であり、出力となるのは閲覧履歴のクラスタである。類似度の算出には termmi⁶⁾ を利用した。termmi は複合語を考慮した単語抽出とベクトル空間法によって閲覧履歴の類似度を算出することができる。

処理のアルゴリズムは図 2 の通りである。

処 理 1

全履歴から以下の条件を満たすものを「強い繋がり」とし、強い繋がりを辿ることでクラスタを形成する。

強い繋がり条件 1 時系列の距離関数：判定する二つの履歴間が一定の閾値 T_1 個の履歴以上離れていない。

強い繋がり条件 2 類似度の関数：判定する二つの履歴間の類似度が一定の閾値 sim_1 を越えている。

強い繋がり条件 3 例外処理^{*1}：判定する二つの履歴間の

*1 戻るボタンで過去のウェブページを経由している場合、経由地となる

Algorithm-process1

Input: a new value $sim(p,q)$,
($p \in \text{all of ID=numid}_1, q \in \text{all of ID=numid}_1$)

Output: $cluster_1$

```

1. for  $x = 1$  to  $numid_1$  do
2.   for  $y = 1$  to  $numid_1$  do
3.     if  $sim(x,y) > sim_1 \cap |x-y| < T_1$  do
4.       Tieconnect $\leftarrow(x,y)$ ;
5.     end if
6.   end for
7. end for
8. foreach Tieconnect(a,b) do
9.   foreach number of cluster do
10.    if  $a \in cluster[num]$  do
11.      cluster[num] $\leftarrow b$ ;
12.    elseif  $b \in cluster[num]$  do
13.      cluster[num] $\leftarrow a$ ;
14.    end if
15.  end foreach
16.  if  $a \in cluster[num]$  do
17.    make new cluster $\leftarrow a, b$ ;
18.  end if
19. end foreach
20. Report (cluster);
```

Algorithm-process2

Input: $cluster_1, sim(p,q), numid_2 = \text{unclusteredID}$

Output: $cluster_2$

```

21. for  $x = 1$  to  $numid_2$  do
22.   foreach  $cluster_1$  do
23.     foreach factor of  $cluster_1$  do
24.       if  $sim(m, \text{factor of } cluster_1) > sim_2 \cap |x-y| < T_2$  do
25.         looseconnectcounter++
26.       end if
27.     end foreach
28.   if factor of cluster*Share < looseconnectcounter do
29.     cluster[num] $\leftarrow numid_2$ ;
30.   end if
31. end foreach
32. end foreach
33. Report (cluster);
```

図 2 要求分類方式のアルゴリズム

類似度が 1 ではない。

なお、分類方式は最短距離法を基に条件を調節したものであり、本処理の処理は強い繋がりを有する全ての閲覧履歴がいずれかのクラスに属するまで繰り返される。たとえば、閲覧履歴 1~6 に対して処理を行い、(1 と 2, 1 と 4, 3 と 6, 4 と 5) の 4 つの強い繋がりを有する場合、(1, 2, 4, 5), (3, 6) の二つのクラスが形成される。

処理 1 で形成された、履歴を要素とするクラスを「クラス 1」とする。つまり、処理 1 が終了した時点で複数のクラス 1(クラス 1-1, クラス 1-2, ... クラス 1-n) が生成されている。

処 理 2

処理 1 で網羅されなかった履歴を対象に以下の条件を満たす「弱い繋がり」を基準に処理 1 で形成されたクラス 1 に処理 1 で網羅されなかった履歴を組み込んでいく。クラス 1 に組み込まれる閲覧履歴の条件とはクラス 1 を構成する要素となる閲覧履歴の一定割合以上に弱い繋がりを有していることである。なお、処理中の閲覧履歴が複数のクラス 1 に対して組み込まれる可能性があるときは、

組み込み可能な全てのクラス 1 に処理中の閲覧履歴を組み込むこととする。

弱い繋がり条件 1 時系列の距離関数：判定する二つの履歴間が一定の閾値 T_2 個の履歴以上離れていない。

弱い繋がり条件 2 類似度の関数：判定する二つの履歴間の類似度が一定の閾値 sim_2 を越えている。

処理 2 で形成された履歴を要素とするクラスを「クラス 2」とする。つまり、処理 2 が終了した時点で複数のクラス 2(クラス 2-1, クラス 2-2, ... クラス 1-m) が生成されている。

処理 1, 処理 2 を経て複数の閲覧履歴クラスが出力される。

3.3 課題解決のアプローチ

本項目では要求分類方式がユーザの要求が生成する閲覧行動にどのようにアプローチしているかを説明する。

これまでの実験の結果、閲覧履歴は以下の 2 つの特徴を有していることが分かった。

閲覧履歴の特徴 1 要求が持続する約 10 分の期間に、平均 20 個程度のウェブページを閲覧しており、その配置は必ずしもクラス重心付近に集中しておらず、樹状のものが多い。

閲覧履歴の特徴 2 また、閲覧履歴のある期間では複数のカテゴリを行き来する形態で混在していた。

閲覧履歴における前者の特徴は研究課題 1 で述べた要求の性質に起因しており、後者の特徴は研究課題 2 で述べた要求の性質に起因していると考えられる。つまり、要求分類方式において、閲覧履歴の特徴 2 点を考慮することで、要求の性質を考慮した分類方式を実現し、分類精度を向上させることができる。

本稿が提案するアルゴリズムは、以下のようにアプローチしている。

閲覧履歴の特徴 1 へのアプローチ

要求ごとに閲覧履歴を分類するためには、樹状のクラスを特定する必要がある。樹上の配置を持つデータを分類するためには、k-means 法などの重心からの距離を利用した融合は適しておらず、局所解を重視した分類が適している。そこで、要求分類方式では局所解を重視した最短距離法による分類を組み入れている。

しかし、最短距離法だけでは精度に関する問題が発生するため、最短距離法で分類困難な閲覧履歴に関しては、類似度を利用した既存クラスへの融合という精度を重視した処理を行う。

閲覧履歴の特徴 2 へのアプローチ

一定期間に異なるクラスの閲覧履歴が混在することを考慮すると、時系列に関する関数を数値化して閲覧履歴間の距離に組み込むことはできない、そのため、時系列に関する関数は出現位置に置き換え、閾値より離れた閲覧履歴同士を強い繋がり、弱い繋がり結び付けないことで、時系列に関する関数が精度を下げるのを抑えた。

また、要求分類方式は誤解析が発生した際、並存する異なるカテゴリの閲覧履歴が連鎖的に同一クラスに融合さ

同じ内容のウェブページの類似度 1 を強い繋がり条件から除くため

せないため、閲覧履歴をクラスタに組み込む前後で融合の基準が変化しない方式 (処理 2) を採用した。

3.4 競合技術への優位性

要求ごとに閲覧履歴を分類する研究はあまり行われていない。そのため本項では、既存の分類方式を閲覧履歴に適用することを想定し、優位性を示す。

提案方式は、ウォード法など一部の既存技術と異なり、各閲覧履歴を表す数値ベクトルを与えるのではなく、閲覧履歴間の類似度を用いて分類を行うアルゴリズムである。既存技術のように因子を与える方が分析に活用できる情報が多いため、分類には適しているが、提案方式は閲覧履歴に留まらず、位置情報、メール送受信履歴、操作履歴といった多様な行動履歴に応用することを目的としている。そのため、距離の逆数という基準で異種行動へ拡張を行いやすい類似度を分類に利用している。

表 1 は以下にあげる競合技術との比較をまとめたものである。

次に代表的な分類方式を紹介し、提案方式の優位性について述べる。

非階層クラスタリング

非階層クラスタリングとは分割と評価関数の再計算を繰り返し、最適な評価値を持つ分割を得る方式である。非階層クラスタリングの代表的な方式である k-means 法を採用した場合、最も大きな問題は分割数をあらかじめ設定しなければならないことである。そこで、ベイズ情報量基準により分割数を自動決定する x-means 法を採用すれば分割数を設定しなくても良いが、情報量基準が正規分布を前提としているため、短期的な要求の抽出に適用するのは難しい。一方、提案方式は短期的な要求に基づいた閲覧行動の性質を前提とするため、より大きい精度を期待できる。

また、k-means 法固有の問題として初期分割に大きな影響を受ける、球形かつほぼ等しい要素数のクラスタに分類することが仮定されている⁹⁾、などが挙げられ、今回対象とする閲覧履歴の分類には適さない。

階層クラスタリング

階層クラスタリングとは近いデータ同士を融合させることで樹形図を作成する方式であり、提案方式の処理 1 では階層クラスタリングの最短距離法の距離算出をもとにアルゴリズムを構築している。

最短距離法とはクラスタ間の距離を計算するとき、最も距離が短くなるデータ同士の組み合わせを採用する方式である。これは、最短距離法を採用した階層クラスタリングは局所解を重視し、データの配置が長い樹状となっているものをまとめるのに最も適しているためである。しかし、最短距離法だけを用いた場合、あらかじめ分割数を設定しなくてはならないという問題や、処理が進むほど精度が下がる (チェイニング効果) という問題が生じる。

特に後者の問題に関しては、最短距離法特有の性質で、クラスタ同士、データとクラスタ、データ同士という組み合わせの順で融合が起こりやすいため、結果として一つの大きな樹状のクラスタを形成する傾向がある。

最短距離法によるクラスタリングにおいて、既に幾つかのクラスタが形成された状態で融合が行われるとき、処理対象データに類似した 1 つのデータがクラスタ内に出現す

表 1 分類方式の比較

	分割数 自動決定	クラスタ の形状	初期分割
非階層クラスタリング			
k-means 法	×	球形に特化	必要
x-means 法	○	球形に特化	必要
階層クラスタリング			
最短距離法	×	樹状に特化	必要
ウォード法	×	球形または樹状	必要
提案方式	○	樹状に特化	不必要

ると、誤ったクラスタに融合されるケースが多発する。この誤融合は処理が進むほど頻繁に起こる。そのため、最短距離法による階層クラスタリングは閲覧履歴の分類においても処理が進むほど精度を下げるることとなる。

提案方式では精度が落ち始める段階で分類を止め類似度ベースの融合を行うため、精度を確保できる。また、処理を切り替える境界となるポイントの閾値 (類似度の値) を設定すれば、クラスタ数を自動決定できる。

階層クラスタリングで最も分類感度が高いとされているのがウォード法である。ウォード法は、各データと属するクラスタの重心の距離を最小化する方式で、対象がベクトルで与えられる必要がある。

しかし、閲覧履歴の分類においては、重心付近で十分なデータ数を確保できず、また、樹状のクラスタを形成することが多いため、クラスタの重心と属する閲覧履歴全ての距離を基準とするウォード法は精度を下げるることとなる。

4. 評価実験

4.1 評価方法

提案方式の分類精度を評価するためにウェブ閲覧行動を対象とした実験を行った。

閾値の設定

実験における閾値はそれぞれ $T_1 = 20$, $T_2 = 20$, $sim_1 = 0.6$, $sim_2 = 0.3$ と設定する。

なお、 $T_1 = 20$, $T_2 = 20$ という閾値は過去の研究における 1 時間に 60 程度のウェブページを閲覧し、毎時 4 回程度の頻度で要求が変化するという知見を根拠として設定した。また、本実験における類似度の分布を考慮し、50% 程度の閲覧履歴を処理 1 の処理対象とするため $sim_1 = 0.6$ とし、ほぼ全ての閲覧履歴がいずれかの閲覧履歴と 0.3 以上の類似度を有しているため $sim_2 = 0.3$ とした。また、70% 以上の閲覧履歴をいずれかのクラスタに属させるため $sim_2 = 0.3$ としたことを考慮し、 $Share = 2/3$ とした。

評価対象となる閲覧履歴の作成

ウェブリテラシーを有した 24~26 才の被験者 5 名 (男性 3 名、女性 2 名) による実験を行なった。被験者は Wikipedia¹⁰⁾ サイト内を閲覧履歴 (日時、タイトル、URL) を取りながら 2 時間 (約 60 履歴分) 巡回し、要求が変化するポイントとなった閲覧履歴をマーキングする。

以上の処理を経て作成した閲覧履歴を利用し、単一要求の閲覧履歴 (以降、要求多重度 1 と呼ぶ)、二つの要求が並存する閲覧履歴 (以降、要求多重度 2 と呼ぶ) の 2 種類の

閲覧履歴^{*1}を作成し、それぞれを評価対象とした分類評価実験を行った。

正解分類の作成

被験者が要求が変化しとしてマーキングを行った閲覧履歴から次のマーキングされた閲覧履歴までを一つのクラスタとし、このクラスタを被験者の要求として正解となるクラスタ群を作成した^{*2}。

比較方式

分類結果を比較する分類方式として階層クラスタリングにおいて高い精度を有するワード法と非階層クラスタリングの中で最も一般的な k-means 法を採用した。両者ともに、正解分類のクラスタに含まれる閲覧履歴群との類似度を各閲覧履歴を表す数値ベクトルとし、分割数には正解データの有するクラスタ数を与えた。また、k-means 法の初期重心はランダムによって決定する。

評価手法：Adjusted Rand Index

実験における評価手法として Adjusted Rand Index¹¹⁾ (以降 ARI) を採用した。ARI とは同一の分類対象を有する二つの分類方式の類似性を図るもので、一方を提案方式による分類結果、一方を正解分類結果として ARI を適用することで分類方式の評価を行うことができる。一般に ARI 値は基本的に 0~1 の値をとり、1 で完全一致、0 でランダムによるクラスタリングの期待値となるが、ランダムクラスタリングの期待値を下回る分類が行われた場合、負の値をとることもある。

評価手法として精度や再現率をとる手法が考えられるが、今回のケースでは正解分類結果と提案方式による分類結果がそれぞれ形成するクラスタ数が異なり、評価方式のクラスタと正解クラスタを関連付ける指標もないため、適用が難しい。また、共通要素が多いクラスタ同士を関連付けて精度を出すとベースラインが 0% にならないという問題がある。

なお、提案方式は全ての閲覧履歴をクラスタに属させる方式ではないため、正解分類からクラスタに属なかった閲覧履歴を除去し、分類が行われた閲覧履歴のみを評価対象として評価を行った。

4.2 実験結果と分析

実験の結果、比較方式の ARI 値が 0 を下回った (ランダムによる分類を下回る分類精度) 被験者のデータについては、被験者によるマーキングが適切でなかったと推測されるため、外れ値とし、提案手法、比較方式の平均 ARI 値算出の対象から除外した。また、提案手法についてはワード法で外れ値としたデータを除去した ARI 平均値を表 2 に示した。なお、ワード法は 1 被験者のデータを、k-means 法は 2 被験者のデータを外れ値としており、図 3 における実験データ数は要求多重度 1 においてワード法 4 個、提案手法 4 個、要求多重度 2 においてワード法 6 個、提案

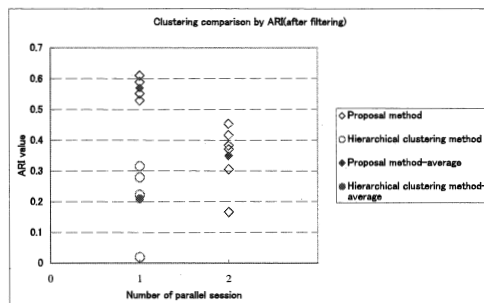


図 3 提案方式とワード法の ARI 値比較

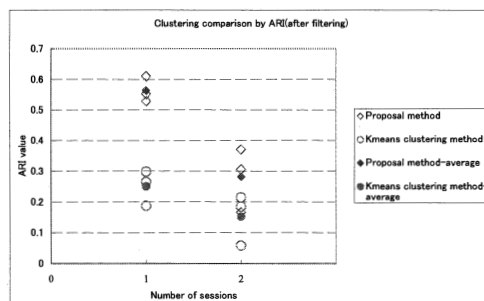


図 4 提案方式と k-means 法の ARI 値比較

手法 6 個となる。図 4 における実験データ数は k-means 法 3 個、提案手法 3 個、要求多重度 2 において k-means 法 3 個、提案手法 3 個となる。

実験データから不適切な閲覧履歴を外れ値として、平均 ARI 値を算出したものが表 2 である。

表 2 平均 ARI 値の比較

	要求多重度 1	要求多重度 2
ワード法	0.208999196	
k-means 法	0.250056395	0.182192223
提案方式	0.569701338	0.349559052

図 3, 図 4, 表 2 が示す通り、要求多重度 2 以下の閲覧履歴分類における平均 ARI 値は提案方式が k-means 法とワード法を上回る結果となった。また、要求多重度 1 の閲覧履歴分類より要求多重度 2 の閲覧履歴分類が困難なため、方式にかかわらず要求多重度 1 における平均 ARI 値は要求多重度 2 における平均 ARI 値より高い数値を示した。

なお、ワード法は 1 被験者のデータを、k-means 法は 2 被験者のデータを外れ値としており、要求多重度にかかわらず、k-means 法は分割数の少ない被験者 (マーキング数が 3 個以下の被験者) の閲覧履歴の分類精度が低い結果となった。しかし、フィルタリング後の ARI 値については k-means 法がワード法をやや上回った。

以上のように、評価実験を通して閲覧履歴分類における提案方式の有効性が示された。

*1 2 つの要求が並存する閲覧履歴は 2 人のユーザの閲覧履歴の開始時刻を合わせ、両者を時系列順にソートすることで混合し、仮想的に作成した。つまり、約 120 履歴で構成される閲覧履歴となる。

*2 たとえば n 番目の要求クラスタは、 $n-1$ 番目にマークされた閲覧履歴の次の閲覧履歴から始まり、 n 番目にマークされた閲覧履歴で終わる。

4.3 考 察

実験を通して得られた知見を以下に示す.

考察 1 提案方式は処理 2 において, 既存クラスタへの振り分けを行っているため, 既存の階層クラスタリングに比べ, 高い分類精度を有している.

考察 2 閾値などの条件をそろえれば, 要求が並存化すると正解クラスタが細分化されるため, ARI 値が下がる.

考察 3 提案方式は全体として過結合の傾向があり, 過剰に長いクラスタが形成される. これは, 処理 1 で最短距離法を基にクラスタリングを行っているため分散が鎖状となる変化を捉えやすい反面, 処理の都合上, データとデータの融合よりクラスタとデータの融合が有利になってしまうためである.

考察 4 正解データは要求変化を被験者自身に示してもらうことで獲得したが, 誤分類の中にも整合性の取れた分類があり, 多様な解が存在する.

考察 5 処理 2 の対象となる閲覧履歴は要求発生直後や要求終了直前に多く発生しており, 要求変化前後においては類似した閲覧履歴が出現しにくい.

5. ま と め

本稿では, 要求の変化構造の把握に向けて, ウェブ閲覧履歴を要求ごとに分類する方式を提案した. さらに, 比較実験を通して, 提案方式が既存方式に比べ高い評価値を有することを確認した.

参 考 文 献

- 1) アマゾン ジャパン株式会社,
<http://www.amazon.co.jp/>.
- 2) 長野, 高橋, 中川, コンテキストを変化させる閲覧履歴の抽出, 人工知能学会第 22 回全国大会, 1F2-8, 2008.
- 3) 山田, 中小路, 上田, インターネットユーザ間の長期にわたる興味遷移パターンの抽出と比較, 第 19 回人工知能学会全国大会論文誌, 2C1-3, 2005.
- 4) 戸田, 福田, 石川, Blog 記事のクラスタリングに基づいたカテゴリ別話題変遷パターンの抽出, データ工学ワークショップ 2007, A8-blog, 2007.
- 5) D. Pelleg and A. Moore, X-means: Extending kmeans with efficient estimation of the number of clusters, ICML2000, Vol17, pp727-734, 2000.
- 6) Windows 用テキストマイニングツール termmi,
<http://gensen.dl.itc.u-tokyo.ac.jp/termmi.html>.
- 7) 中川, 森, 湯本, 出現頻度と連接頻度に基づく専門語抽出, 自然言語処理, Vol10, No1, pp27-45, 2003.
- 8) M.Salton, M.J.McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1983.
- 9) S.Guha, R.Rastogi, K.Shim, CURE: An Efficient Clustering Algorithm for Large Databases, Proceeding of the ACM SIGMOD International Conference on Management of Data, pp73-80, 1998.
- 10) Wikipedia, <http://ja.wikipedia.org/wiki/>.
- 11) Hubert, L. and Arabie, P, Comparing partitions. Journal of Classification, pp193-218, 1985.