ウェブ閲覧履歴に反映される要求変化の抽出方式の提案

長野 翔一*1 高橋 寛幸 中川 哲也

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

ウェブ閲覧履歴からユーザの興味を推測するプロファイル技術は，想定しているよりも短い周期で変化する閲覧行動を対象とする場合，分析に必要な閲覧履歴数が確保できないため，適用が困難である．本研究においては 10 分程度の短い周期で出現する行動への動機を「要求」と定義し，閲覧履歴に反映されるユーザの要求変化を捉えることが可能なプロファイル技術の実現を目指す．

そこで，我々は閲覧履歴におけるテキスト間の類似性と出現位置を用いて，閲覧履歴を生成した要求ごとに分類する要求分類方式と，分類されたクラスタ群からクラスタ間の変化関係を抽出する関係抽出方式を提案する．

また，既存の分類方式との比較実験を通して，要求分類方式が既存方式に比べ，より被験者の入力に近い分類を行うことを確認した．

Clustering and structurizing the access-log for detecting dynamic intentions

Shouichi Nagano, Hiroyuki Takahashi, Tetsuya Nakagawa

NTT Information Sharing Platform Laboratories, NTT Corporation

We propose a clustering and structurizing methods for treating the change of intention from user's browsing behavior.

It is necessary to treat user's intention accurately in information explosion. However, treating dynamic intention is difficult for a conventional method, as a behavioral targeting method.

For detecting user's intention context in access-log, we analyze each of browsing-history based on the similarities of meaning, so that clustering and structurizing methods visualize intention change from access-log.

We report on result of an experiment to effectiveness for conventional clustering method. Additionally, for relating clusters, we extract an access-log which prompt to change intentions, and report on visualized result of experiment.

1. はじめに

近年，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，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) から，個々のユーザの興味を把握するプロファイル技術が不可欠である．

「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)[2]，戸田 (2007)[3] 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[4] 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，頻繁に変化しない性質を持つ興味を対象としており，閲覧履歴を時系列順に取得し，獲得した閲覧履歴全体からユーザの閲覧行動から興味を推測する方式をとる．そのため，プロファイル技術が頻繁に変化する性質を持つ要求を対象とする場合，要求の変化を検出することが不可欠となる．

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，要求によって生成された閲覧履歴を，要求ごとに分類する要求分類方式と，分類されたクラスタ同士の関係を抽出する関係抽出方式を提案する．

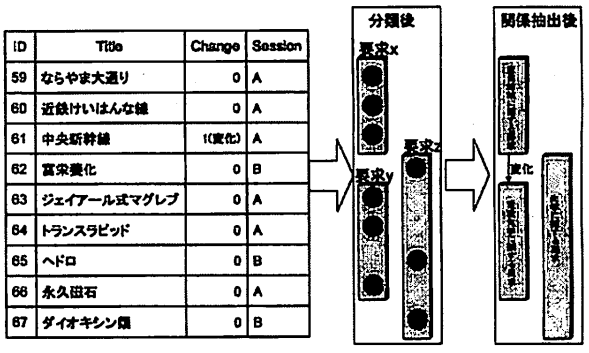

図 1 評価対象となる閲覧履歴の一例

提案する要求分類方式と関係抽出方式について，以下に具体例を示す．

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．

閲覧履歴タイトルからこのユーザの要求を推測すると，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と考え

*1 〒 180-8585 東京都武蔵野市緑町 3-9-11
Tel 0422-59-3397
Fax 0422-59-5657
Mail nagano.shouichi@lab.ntt.co.jp

られる．

この閲覧履歴を分類，関係抽出する場合，以下のような処理が行われる．はじめに，閲覧履歴に対して本稿で提案する分類方式を適用し，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を行う．

次に，ID.61 の中央新幹線がユーザの奈良地域に関する要求を電磁気学に関する要求に変化することを促していると考えられるため，変化元となる閲覧履歴群 (要求 x から生成された閲覧履歴) と変化先となる閲覧履歴群 (要求 y から生成された閲覧履歴) を変化関係として紐付ける．

以上のように，本稿では分類された閲覧履歴とそれらの変化関係を紐付けたデータを獲得する．

このデータを利用することで，要求変化が多いユーザを対象としたプロファイル技術の精度向上が期待され，また，ある時点におけるユーザ要求がどのような変遷を経ているかという背景情報を獲得すること可能となる．

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 章において要求分類方式を説明し，競合への優位性，研究の位置付けについて示す．4 章において要求分類方式の有効性を検証した評価実験について示す．5 章において分類された閲覧履歴からクラスタ同士の関係を抽出する関係抽出方式の提案について示す．最後に 6 章においてまとめについて示す．

2. 背　　景

2.1 閲覧履歴の分類と関係抽出の必要性

閲覧履歴からユーザの興味を推測するプロファイル技術はユーザの興味が変化しない，または緩やかに変化することを想定している．これまで，プロファイル技術は閲覧履歴を期間ごとに分割することで，要求変化への対応を試みてきた．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，プロファイルの精度が下がる．

そこで，本稿では要求分類方式，関係抽出方式を実現することで，要求が頻繁に変化するユーザの閲覧履歴を対象とした要求を推定可能とし，また，ある時点における要求がどのような変化を経ているのかという背景情報の獲得を目指す．

2.2 閲覧履歴から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，約 10 分という短期間しか持続しない要求において，閲覧履歴の分析は閲覧履歴数が少なくなるため，困難であった．

要求の変化を捉えるための既存方式として，行動ターゲティング広告などで利用される興味プロファイルの重み付け技術がある．これは，閲覧履歴を一定期間ごとに分割し，プロファイルの重みを変化させる方式である．しかし，分割された期間内に要求変化が起こっている場合，要求変化前の閲覧履歴がノイズとなり，プロファイル構築の精度を下げることとなる．

また，閲覧履歴を分類し，クラスタの特徴値から興味遷移を捉える研究も行われている．例えば，山田 (2005)[2] は興味遷移を捉えるため，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，閲覧履歴数が少ない要求変化においては，分類したクラスタの要素数が減少するため，クラスタ特徴値の変化を利用して，変化を捉えることは困難である．

我々は，分割に関する問題を解決するため，ユーザの要求ごとに分類し，クラスタの関係をたどることで要求の変化を捉える．また，分類における履歴数確保の問題を解決するため，要求の性質が閲覧履歴の特徴として表出することを考慮し，閲覧履歴の性質を利用した分類を行う．

3. 閲覧履歴の分類方式の提案

3.1 要求分類方式の提案概要

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，分析が困難となっていると考える．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 要求分類方式のアルゴリズム

本項では，要求分類方式のアルゴリズムについて述べる．要求分類方式は処理は 2 段階に分けて行われ，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う．処理のアルゴリズムは図 2 に示す．なお，処理で用いられる閾値は，次のような目的で設定する T_1, T_2 は出現場所の制限のため設定する履歴間の距離の閾値である．sim_1 は処理 1 で対象とする閲覧履歴の絞込みのため設定する類似度の閾値である．sim_2 は処理 2 で対象とする閲覧履歴の絞込みのため設定する類似度の閾値である．$Share$ は処理 2 で対象とする閲覧履歴の絞込みのため設定するクラスタ内における閲覧履歴の割合の閾値である．

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．

類似度とは，2 つの文書の内容がどの程度類似しているかを示す尺度である．類似度の算出法は様々な手法が提案されているが，今回，類似度の算出には termmi[6] 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を利用し，2 つの文書を構成する単語群の類似性を数値化することが可能である．

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類似度が 1 ではない．

なお，要求分類方式は高類似度の閲覧履歴同士に対して最短距離法による融合を行っており，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タ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) の二つのクラスタが形成される．

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

---

*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

```
Algorithm-process1
Input:  a new value sim(p.q),
(p ∈all of ID=numid1, q ∈ all of ID=numid1)
Output: cluster1
1.  for x = 1 to numid1 do
2.    for y = 1 to numid1 do
3.      if sim(x,y) > sim1∩ | x − y |< T1 do
4.        Tieconnect←(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 ∈ cluster[num] do
11.       cluster[num]←b;
12.     elseif b ∈ cluster[num] do
13.       cluster[num]←a;
14.     end if
15.   end foreach
16.   if a ∈ cluster[num] do
17.     make new cluster←a,b ;
18.   end if
19. end foreach
20. Report (cluster);
Algorithm-process2
Input:  cluster1, sim(p.q), numid2 =unclustedID
Output: cluster2
21. for x = 1 to numid2 do
22.   foreach cluster1 do
23.     foreach factor of cluster1 do
24.       if sim(m,factor of cluster1) > sim2∩ | x − y |< T2do
25.         looseconnectcounter++
26.       end if
27.     end foreach
28.     if factor of cluster*Share < looseconnectcounter do
29.       cluster[num]←numid2;
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

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タ 1 に処理 1 で網羅されなかった履歴を組み込んでいく．クラスタ 1 に組み込まれる閲覧履歴の条件とはクラスタ 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タ 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タ 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タを「クラスタ 2」とする．つまり，処理 2 が終了した時点で複数のクラスタ 2(クラスタ 2-1，クラスタ 2-2，$\cdots$ クラスタ 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タが出力される．

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タ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，樹状のクラスタ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タ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タ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，弱い繋がりで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タに融合させないため，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．そのため本稿では，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．

要求分類方式は，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，要求分類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．

以下にに代表的なクラスタリング方式を紹介し，要求分類方式との詳細な比較について述べる．

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，要求分類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，要求分類方式の処理 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

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

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こととなる．

要求分類方式では精度が落ちはじめる段階でクラスタリング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こととなる．

4. 要求分類方式の評価実験

4.1 実験方式

要求分類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 Wikipedia[10] 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) 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た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，このクラスタをユーザの要求として正解となるクラスタ群を作成した*2．

比較手法

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ウォ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テキストと各データ (閲覧履歴) の類似度を因子として，分割数として正解データの有するクラスタ数を分割数として与えた．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 Index[11] (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，一方を提案手法による分類結果，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評価手法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%とならないという問題がある．

なお，要求分類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さ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ウォ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，ウォード法は 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 においてウォード法 4 個，提案手法 4 個，要求多重度 2 においてウォード法 6 個，提案手法 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，図 3，図 4 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 値は要求分類方式が k-means 法とウォ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

表 2 ARI 値の比較

	要求多重度 1	要求多重度 2
ウォード法	0.208999196	0.160903986
k-means 法	0.250056395	0.182192223
提案方式	0.569701338	0.349559052

*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．

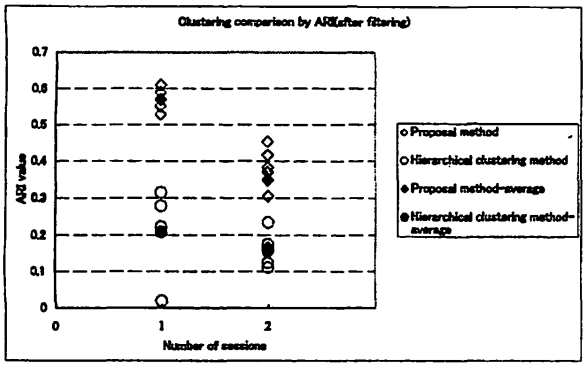

図 3 要求分類方式とウォード法の ARI 値比較

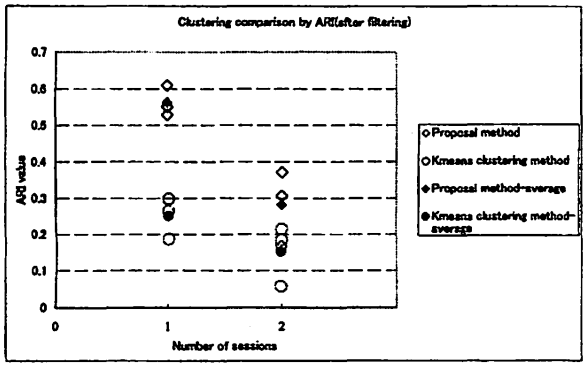

図 4 要求分類方式と k-means 法の ARI 値比較

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，ウォード法は 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 法がウォ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要求分類方式の有効性が示された．

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 要求分類方式は処理 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 要求分類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分類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と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ユーザ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. 関係抽出方式の提案

5.1 特異点を利用したクラスタの関係抽出方式の概要

本章では 3 章のアルゴリズムで得られた閲覧履歴のクラスタから変化関係を抽出し，クラスタ同士を紐付ける関係抽出方式を提案する．

我々はこれまでに，ユーザの要求変化を促進させる閲覧履歴が存在すると仮定し，その存在を検証した[4]．本稿では，その知見を活かし要求変化を促進させる閲覧履歴を特異点と定義した．例えば，図 1 の閲覧履歴を有するユーザは奈良地域に関する要求を電磁気学に関する要求に変化させているが，特異点の存在を仮定すると ID.61 中央新幹線が特異点となり，要求変化を促進させていることとなる．

閲覧履歴において，特異点とは変化元となるクラスタに属する 1 つの要素であり，変化元となるクラスタ内において，変化先となるクラスタとの類似度が相対的に高いものをさす．

そこで，我々は変化元となるクラスタ内において，変化先クラスタと相対的に高い類似度を有する閲覧履歴を特異点として抽出することで，変化元クラスタと変化先クラスタを紐付けることが可能であると考える．

5.2 クラスタの関係を抽出する既存方式の問題点

クラスタ同士の変化を特定する既存方式として，ウェブページの遷移確率を利用し[12]，変化元ページの最も新しい閲覧履歴と変化先ページの最も古い閲覧履歴の遷移確率から変化を抽出する方式が考えられる．

しかし，今回のケースでは十分な学習データが用意できず，また，新しく出現したコンテンツに対応できないという問題がある．

関係抽出方式では，変化先クラスタと特異点の相対的な類似度の高さと，出現位置から紐付けを行う．そのため，新しく出現したデータにも対応可能で，学習を行う必要がない．

5.3 関係抽出方式の詳細

本項目では，クラスタ同士の類似度と出現時期から特異点を特定する方式の詳細を説明する．

特異点の特定は，任意のクラスタを変化先の要求から生成されたクラスタ (以後，変化先クラスタと呼ぶ) として固定し，下記条件に合致した特異点を抽出する．

なお，変化元クラスタが含む，最も新しい閲覧履歴の ID を ID_1 と置く．変化先クラスタが含む，最も古い閲覧履歴の ID を ID_2，最も新しい閲覧履歴の ID を ID_3 と置く．特異点の ID を ID_4 と置く．変化元クラスタが含む全ての閲覧履歴と変化先クラスタが含む全ての閲覧履歴の類似度を p と置く．特異点となる閲覧履歴と変化先クラスタが含む全ての閲覧履歴の類似度を q と置く．sim_3, sim_4, T_3 は閾値とする．

特異点条件 1: $\frac{(ID_2+ID_3)}{2} > ID_1 \cap \mid ID_2 - ID_4 \mid < T_3$

特異点条件 2: $q/p > sim_3 \cap p > sim_4$

特異点条件 3: 条件 1，条件 2 を満たす閲覧履歴が複数存在する場合，q/p の値が最も大きいものを特異点として採用する．また，q/p の値が最も大きいものが複数存在する場合，ID_4 の値が大きいものを採用する．

特異点条件 1 は変化元クラスタ，変化先クラスタ，特異点の出現位置を限定し，特異点を絞り込んでいる．これは，変化先クラスタの中心点までには変化元クラスタが開始しており，変化先クラスタの最も古い閲覧履歴付近に特異点が出現するというこれまでの実験による知見に基づく．特異点条件 2 は特異点と変化先クラスタの相対的な類似度と特異点と変化先クラスタの絶対的な類似度から特異点を絞り込んでいる．q/p とはこれまでの実験で 70%程度の特異点が 1 以上の数値を取った，特異点と変化先クラスタの相対的な類似性を表すパラメータである．特異点条件 3 は特異点条件 1，2 を満たした特異点候補の中から，それぞれの特異点候補と変化先クラスタの相対的な類似度を比較して，1 つの特異点を特定している．

特異点条件により，抽出された特異点が属するクラスタを変化元クラスタとして，変化元クラスタと変化先クラスタを紐付ける．全てのクラスタを変化先クラスタとして特異点が抽出できた時点で処理を終了する．条件に合致する閲覧履歴が存在しないときは変化元なしのクラスタ (新しく発生した要求である) とした．なお，類似度の算出法に

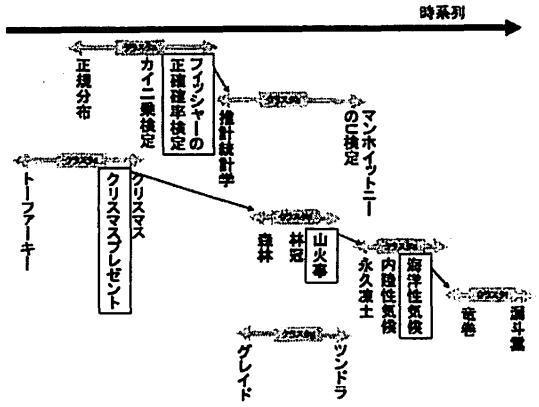

図 5 関係抽出方式による要求の変化関係の可視化

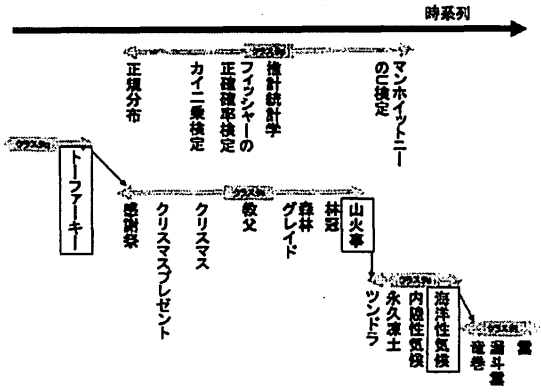

図 6 被験者の入力した要求の変化関係の可視化

ついては 3 章に準ずる.

5.4 関係抽出方式の実験方式

5.3 項の条件により要求の変化関係を紐付け，閲覧者の支援で入力された正解データと比較を行う.

対象データの選定

誤分類の影響を抑制するため，4 章 評価実験の要求多重度 2 において最も高い ARI 値を示した閲覧履歴の一部 (135 履歴分) を実験対象に採用する.

正解データの作成

4 章の実験で使用した正解データにおける，マーキングの前後にあるクラスタを要求変化として紐付けることで作成する.

閾値の設定

本実験における閾値は以下のように設定する.

$T_3 = 20, sim_3 = 1, sim_4 = 0.005$

出現位置に関する閾値 T_3 は要求分類における閾値同様，1 クラスタの平均的な要素数から設定した. sim_3 はこれまでの実験において，q/p が 1 を下回る要求変化が 30%以下であるという知見から設定した. sim_4 は被験者が入力した要求変化における p の最低値を取り，設定した.

これらの数値は，実験対象のデータにあわせて設定しているため，今後の追加実験により適宜調整する.

5.5 実験結果と考察

関係抽出方式による実験結果を受けて，一部 (約 50 履歴分) を可視化したものが図 5，閲覧者の支援で入力された正解データは図 6 の通りである.

中央に「クラスタ」という表記を含む両矢印は，要求分類方式により生成された各クラスタの出現期間を表している. 縦文字は，クラスタに含まれる閲覧履歴のタイトルを表している. 長方形の中のタイトルは，関係抽出方式により抽出された特異点を表している. 特異点から発する矢印は，その特異点によって変化したクラスタへ向かっている.

要求分類方式により過分割されたクラスタは特異点によって紐付けられやすく，過結合 (正解分類より大きなクラスタを形成する誤解析) されたクラスタは特異点条件 p が正しい値を取らないため，誤った特異点抽出の原因となることが多かった. そのため，要求分類方式で形成されるクラスタ数は多いほうが紐付けの精度が高いという知見を得た.

要求分類方式によるクラスタと被験者の入力したクラスタを対応させたとき，関係抽出方式により検出した要求変化 8 個のうち，非正解は 4 個であったが，過分割 (正解分類より小さなクラスタを形成する誤解析) されたクラスタ間を紐付けているものや，正解データ以外の変化が，今後，定量的な評価を行う際には，評価方法を検討する必要がある.

また，被験者の入力した 5 個の要求変化のうち 4 個は抽出されており，要求変化の 80%を抽出できたことから，特異点を利用した関係抽出方式が要求変化を紐付けるために有用であるという見通しを得た.

6. ま と め

本稿ではウェブ閲覧履歴を要求ごとに分類する要求分類方式を提案し，評価実験を通してその有効性を示した.

また，特異点という概念を提案し，特異点による関係抽出方式の構築に取り組んだ. さらに，実験結果を可視化し，正解データの再現性から，関係抽出方式実現の見通しを得た.

謝 辞

本研究を進めるにあたり，アルゴリズムや評価方法に種々の助言をいただいた NTT サイバーソリューション研究所 別所克人研究主任，NTT コミュニケーション科学基礎研究所 平博順博士に謝意を表します.

参 考 文 献

1) アマゾン ジャパン株式会社, http://www.amazon.co.jp/.
2) 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.
3) 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ンの抽出，データ工学ワークショップ 2007，A8-blog，2007.
4) 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.
5) D. Pelleg, A. Moore, X-means:Extending kmeans with efficient estimation of the number of clusters, ICML 2000, 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，Vol10No.1，pp27-45，2003.
8) M.Salton, M.J.McGill, Introduction to Modern Imformation 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, pp.73-80, 1998.
10) Wikipedia, http://ja.wikipedia.org/wiki/.
11) L.Hubert, P.Arabie, Comparing partitions. Journal of Classification, p193–218, 1985.
12) 向，成，上林，利用履歴に基づく PageRank アルゴリズムの改良,DEWS2002，A1-2，2002.