Vol.2014-UBI-42 No.5
2014/5/29

閲覧履歴の関連性を考慮した閲覧意図の階層的分類手法

富士谷 康[1,a)]　吉田 拓磨[1,†1]　中村 明順[1]　安積 卓也[1,†2]　望月 祐洋[1]　西尾 信彦[1,b)]

概要：Web 上での活動支援のためにユーザ個人の閲覧履歴を分類する研究が行われている．分類の際にはユーザの閲覧意図を考慮する必要があるが，時刻やページ本文を用いた既存手法では複数の意図が混在したり，本文が存在しないページの分類ができない．本論文では新たな意図の出現を捉え，閲覧意図の階層性を反映した閲覧履歴の分類を目指し，2 段階の分類手法を提案する．第 1 段階では閲覧の起点とタブ切替に注目した履歴間の関連性を用い機械的に分類する．第 2 段階ではまとめた履歴に含まれる本文の類似度によって階層的クラスタリングを適用することで，分類結果が表す閲覧意図の粒度操作を可能にする．評価では，被験者が意図ごとにまとめた履歴と提案手法による分類結果の一致性が，本文のみによる比較手法より幅広い閾値で高く，閲覧の起点で意図の出現を一定程度の正確性と網羅性で捉えられることを示した．

キーワード：クラスタリング，閲覧履歴，閲覧意図，クリックストリーム，ウェブ

Hierarchical Clustering Method for User Browsing Intentions based on the Relationship in Browsing History

Abstract: Many researches have been performed on clustering methods for personal web browsing history. Existing clustering methods using the visit time and/or text contents cannot reflect user's intentions. This paper proposes a two-phased clustering method suited for capturing the appearance of a user's new intention along with reflecting the hierarchical structure. In the first phase, we create groups of history by applying a clustering method based on the relationship in browsing history. In the second phase, we apply a hierarchical clustering method using the similarity of text contents in order to control the granularity of an intention. The conformance rate was evaluated between the results grouped manually by research participants and grouped automatically by proposed method. The results show the effectiveness of proposed method compared with a method only using a document clustering. Moreover, proposed method can capture the appearance of intentions in a precise and comprehensive manner.

Keywords: clustering, browsing history, browsing intention, clickstream, web

1. はじめに

Web 上のページ数の増加に伴って，個人に適した情報の取得が困難になりつつあることを背景に，パーソナライゼーションが注目されている．情報推薦や検索支援といった Web 上での活動支援を実現するために，個人の閲覧履歴を分類する研究が行われている [1], [2]．分類結果は，ユーザプロファイルの構築や既読情報を再検索するリファインディング [3] 支援などに利用される．閲覧中の興味や目的を正確に捉えるため，閲覧履歴の分類はユーザ自身が考える閲覧意図を捉えていることが望ましい．既存手法として，閲覧時刻を利用する手法では，複数の興味や目的を満たすために様々なページを並列的に閲覧した場合，分類結果に複数の閲覧意図が混在する．ページ本文を利用する手法では，PDF や画像といった本文を抽出できないページの分類ができない．ページ遷移軌跡を表すクリックストリームは，パーソナライゼーションやユーザビリティ向上のために，サイト利用者の行動分析などで利用される [4]．これを用いることで，推移する閲覧意図をある程度追従できると考えられるが，遷移関係が存在しない場合，細かく分類され，ユーザが考える閲覧意図に合わない可能性がある．

[1] 立命館大学 Ritsumeikan University
[†1] 現在，株式会社野村総合研究所
Presently with Nomura Research Institutee, Ltd.
[†2] 現在，大阪大学 Presently with Osaka University
[a)] fujiya@ubi.cs.ritsumei.ac.jp
[b)] nishio@cs.ritsumei.ac.jp

本論文では，個人の閲覧履歴を対象に，新たな意図の出現を捉え，閲覧意図の階層性を反映した分類を目指す．提案手法では閲覧履歴を 2 段階で分類する．第 1 段階では，閲覧履歴に含めたページ遷移やタブに関する情報を用いて，閲覧履歴の関連性に注目し，ページの内容解析や閲覧時刻を利用せず機械的にまとめることで，閲覧意図を最細粒度で表すと期待されるまとまり（マイクロクラスタ）を構築する（これをマイクロクラスタの構築と呼ぶ）．マイクロクラスタにおいて過去の閲覧履歴と関連性が存在しない履歴（閲覧の起点）を，意図の出現とみなし，これを獲得する．第 2 段階で，マイクロクラスタに含まれる本文を用い，その類似度によって階層的クラスタリングを適用することで，閲覧意図が共通するマイクロクラスタを階層的に併合する（これをマイクロクラスタの階層的併合と呼ぶ）．併合結果に任意の閾値を指定することで，最終的な閲覧履歴のまとまり（最終クラスタ）を得る．これによって最終クラスタが表す閲覧意図の粒度操作を可能にする．

本論文の構成は次のとおりである．2 章で本論文における閲覧意図について述べる．3 章で本論文の関連研究を挙げ，4 章で提案手法について述べる．5 章で提案手法の評価を示し，6 章でまとめと今後の課題を述べる．

2. 閲覧意図

興味や目的を満たすために行うページの閲覧には，ユーザ自身が考える閲覧に至った意図（閲覧意図）を伴う．閲覧意図には個人差があるが，例えば，「Java についての検索を行う」「メールを作成する」といったことである．

閲覧意図は Web 閲覧中に推移することが指摘されている．長野ら [5] によれば，閲覧行動への動機を指す「要求」は 10 分程度で変化する性質を持つとしている．しかし実際には，ブラウザのタブを利用することで，複数の閲覧意図を満たすような並列的な閲覧ができる．我々が扱う閲覧意図は，タブ操作を考慮し，頻繁に変化すると想定している．閲覧意図の推移を考慮し，新たな意図の出現を網羅的かつ正確に獲得できる分類は，複数の意図が混在しにくく，適切なタイミングでの活動支援へ応用できると考えられる．

閲覧意図は，階層構造を持つ傾向がある．例えば，大まかに「Java」について調べている時，その内部には「Javadoc についての検索」や「Java クラスについての検索」といった細かな意図が含まれる．この階層構造において，閲覧意図の単位あたりの大きさは，閲覧意図の粒度として考えることができる．閲覧意図の「粗い」「細かい」といった粒度は，ユーザ自身の考え方や分類結果の利用目的などに依存するため，あらゆるユーザが適切とする普遍的な正解や，間隔尺度を持つような絶対的・客観的な正解はなく，主観的で相対的な，順序関係のみを持つと考えられる．これは例えば，「Java に関する閲覧」をより粗く「プログラミングについての閲覧」と考えることもできるため，どの程度を粗い意図とするのか，または細かい意図をどの程度細かくするのか，といったことには個人差がある．

本論文では，意図の出現と階層構造を明らかにすることで，粒度操作ができる分類を目指す．適切とする粒度には個人差があるため，ある特定の粒度に合わせて機械的に分類する意味はなく，閲覧意図の階層性を反映した分類によって，最終クラスタが表す閲覧意図の粒度を，必要に応じて容易かつ自由に操作できる必要がある．粒度操作を可能にすることで，例えば，リファインディング支援に応用すれば，はじめに粗い分類結果を提示した後，ユーザが想起したい情報へのズーミングができる．他にも，粗い意図に合う最終クラスタは嗜好の抽出に，一方，細かい意図に合うものは閲覧した時々における興味の抽出への利用が期待できる．

3. 関連研究

長野ら [1] は，短期的な興味プロファイルの構築を目指し，閲覧履歴が少ない場合でも有効な分類手法を提案している．しかし，この手法では，ページ本文のみから閲覧履歴間の類似度を定義しているため，PDF や画像といった本文が抽出できないページの分類を行えない．

飯野ら [2] は，リファインディング支援のために閲覧履歴を階層的に分類している．分類の指標として，頻出名詞の一致率，URL ドメインの一致・不一致，閲覧時刻の近さ，遷移元と遷移先の関係を用いている．しかし，検索結果ページなどの汎用性の高いページのドメインや閲覧時刻の近さを利用しているため，分類結果に複数の閲覧意図が混在してしまい，閲覧意図を捉えられない．

丸山ら [6] は，検索支援を目的として，ページ遷移およびタブの切替えを考慮したクリックストリームグラフを構築し，閲覧時間を元に検索時における重要なページを抽出している．これによって，検索時の行動の起点や終点をある程度見極められるとしているが，検索以外の閲覧を考慮していない．さらに，クリックストリームのみでは，閲覧意図が継続するものの遷移関係が存在しない場合には分類できず，細かな意図しか捉えられない．

4. 提案手法

本論文では，閲覧履歴を 2 段階で分類する手法を提案する．提案手法の概要を**図 1** に示す．マイクロクラスタの構築では，閲覧履歴に含めた情報を利用し，閲覧履歴の関連性に着目することで，ページ内容や閲覧時刻を利用せず，機械的に分類する．本文を抽出できないページの閲覧履歴を他と連結することで，この分類が可能になることや，閲覧時刻を利用しないことで閲覧意図の混在を防ぐことが期待できる．この段階で閲覧意図を最細粒度で網羅的に獲得することを目指す．マイクロクラスタの階層的併合では，本文の類似度による階層的クラスタリングを適用すること

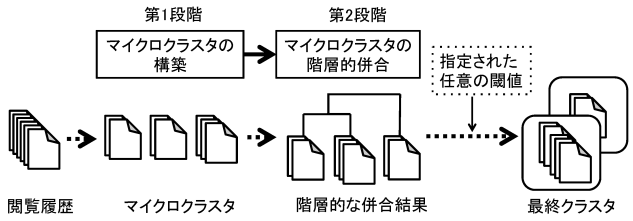

図 1　提案手法の概要
Fig. 1　Schematic of proposed method.

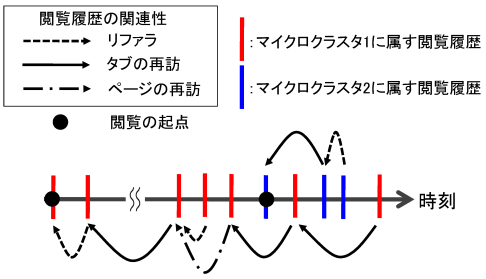

図 2　閲覧履歴の関連性によるマイクロクラスタの構築
Fig. 2　Construction of micro-cluster with the relationship in browsing history.

で，閲覧意図の共通部分を持ったマイクロクラスタ同士を階層的に併合する．併合結果に対して任意の閾値を指定し，最終クラスタを獲得する．これによって閲覧意図の階層構造を捉え，粒度操作を可能にする．はじめに分類対象とする閲覧履歴を定義し，続いて，提案手法について述べる．

4.1　閲覧履歴

本論文で対象とする閲覧履歴は，ページへの訪問履歴に加え，タブを切替えてブラウザの前面にした履歴（タブ前面履歴）を合わせたものとする．訪問履歴のみではタブを用いた閲覧を追従できないため，タブ前面履歴を加えている．閲覧履歴には，ページの URL や，ソーステキスト，ブックマークやリンクによる遷移であるかといった遷移方法や遷移元の情報，およびページを開いたタブの識別子といったタブに関する情報などを含めた．

4.2　マイクロクラスタの構築

我々は，閲覧の起点と関連性によって閲覧履歴を分類する手法を提案している [7]．閲覧履歴の関連性は，過去の閲覧時と閲覧意図が変化しないと考えられる閲覧行動に着目して，リファラ（遷移元の閲覧履歴情報）の存在，タブの切替えによるすでに開いていたタブへの再訪（タブの再訪），「戻る」「進む」などによる既訪のページへの再訪（ページの再訪）の 3 つと定義する．

関連性が存在する閲覧履歴同士を連結することで，マイクロクラスタを機械的に構築する．この概要を**図 2** に示す．マイクロクラスタにおいて，過去の閲覧履歴と関連性が存在しない閲覧履歴を閲覧の起点とする．マイクロクラスタでは閲覧意図が変化しない閲覧行動に着目していることから，最細粒度の閲覧意図を捉えることが期待され，これができれば，閲覧の起点を意図の出現とみなせる．

4.2.1　リファラ

リンクをクリックしてページを遷移した場合などには，遷移先の閲覧履歴にリファラが付与される．リファラが付与される閲覧では，遷移元を閲覧した時と閲覧意図が変化しない可能性が高く，関連性があるとする．リンクによる遷移時に，リダイレクトや，HTTPS のページから HTTP のページへの遷移が行われた場合，リファラが付与されないことがある．これに対処するため，閲覧履歴がリンクによる遷移であり，かつ直前の閲覧履歴とドメインが一致する場合，その 2 つに関連性があるとする．これによって，意図の出現をより正確に捉えられるようになる．

一方，リファラが存在していても，閲覧意図が変化する場合もある．例えば，ユーザが「Java」について検索を行い，検索結果ページを閲覧する．その後閲覧意図が変わり，クエリを「Java」から「旅行」に入れ替えて検索した場合，「旅行」の検索結果ページは「Java」の検索結果ページのリファラを持っているため，直前のクエリに関係なくそれらの閲覧履歴が連結されてしまう．そこで，同一の検索エンジン [*1] において検索結果ページのリファラが別の検索結果ページの閲覧履歴を指す場合，URL からクエリを取得する．クエリが部分一致しない場合には閲覧意図が変化したと判断し，閲覧の起点とすることで，検索を繰り返す際の意図の出現を捉える．この操作をクエリ不一致による起点の生成と呼ぶ．

4.2.2　タブの再訪

タブへ再訪した際に発生する閲覧履歴は，直前の同一タブの閲覧と閲覧意図が変化しないと考え，閲覧履歴間に関連性があるとする．タブの再訪時に閲覧意図が変化しない閲覧とは．例えば，「旅行」についてのページをあるタブで開き，別のタブで「Java」について調べた後，再度「旅行」のタブを開く場合，「旅行」に関する閲覧を指す．

4.2.3　ページの再訪

「戻る」や「進む」などによって発生した閲覧履歴と，直前の同一タブでの同一ページの閲覧履歴間には関連性があるとする．例えば，あるタブでページ A に訪れた（履歴 a）後，同一タブでページ B に訪れ（履歴 b），「戻る」を利用してページ A に再訪する（履歴 c）．この時，履歴 c は，履歴 a と関連性を持つ．ページの再訪はページ遷移を伴う閲覧行動であり，タブの再訪と異なる閲覧行動によって発生するため，これらを区別している．

4.3　マイクロクラスタの階層的併合

マイクロクラスタの階層的併合では，マイクロクラスタ

[*1] 本論文では Google 検索のみを対象としている

に含まれる本文を用い，階層的クラスタリングを適用する．併合結果には閲覧の起点が保持されており，任意の閾値を指定し最終クラスタを構築する．閾値の操作によって，閲覧意図の粒度を自由に操作ができる分類を実現する．この操作は，ユーザとの直接的，もしくはアプリを介した間接的な対話によって行うことを想定している．閲覧意図の階層構造を捉えられれば，ある閾値で作成した最終クラスタを 1 つユーザに提示し，「より粗く」「より細かく」というように求める粒度を明示的に示してもらうことで，ユーザ自身が考える粒度に合った分類結果を提示できる．

続いて，階層的クラスタリングの処理の流れについてを述べる．はじめに，閲覧履歴のソーステキストからタグなどを取り除きページ本文を抽出する．ベクトル空間法を用いるため，閲覧履歴の本文を結合し，マイクロクラスタごとに文書を構築する．その後，文書に対して MeCab[*2] を用いて形態素解析を行い，品詞に分解する．文書に含まれる名詞のみを特徴語として，TF-IDF 手法で重み付けを行い，各文書の特徴ベクトルを構築する．文書 k に含まれる特徴語 t_{ik} を式（1）で示すように重み付ける．tf_{ik} は文書 k における特徴語 i の出現頻度である．式（2）の N は文書総数を表し，df_i は特徴語 i が出現する文書の数を表す．マイクロクラスタの構築によって，文書ごとに文書長が大きく異なることがあるため，式（3）に示すコサイン正規化を行う．式中の m は特徴語の種類数を表す．

$$t_{ik} = \frac{tf_{ik} \cdot idf_i}{n_i} \tag{1}$$

$$idf_i = \log\left(\frac{N}{df_i}\right) + 1 \tag{2}$$

$$n_i = \sqrt{\sum_{i=1}^{m} \left(tf_{ik} \cdot idf_i\right)^2} \tag{3}$$

式（4）に示す文書間距離には，式（5）のコサイン類似度を用いた．今回，階層的クラスタリングの手法のうち，群平均法を適用した．

$$distance\,(x, y) = \frac{2 \cdot cos^{-1}\,(cos\,(\theta))}{\pi} \tag{4}$$

$$cos\,(\theta) = \frac{x \cdot y}{\parallel x \parallel \parallel y \parallel} \tag{5}$$

階層的クラスタリングの結果に対して任意の閾値 $threshold$ を与えることで最終クラスタを形成する．$threshold = 0$ において作成される最終クラスタはマイクロクラスタと等しい．

5. 実装と実験

本章では，提案手法の実装について述べた後，評価実験の詳細や評価結果を述べる．評価は，被験者自身に意図ごとにまとめてもらった閲覧履歴と提案手法による分類結果との一致性に加え，閲覧の起点が意図の出現を網羅的かつ正確に捉えられているかの観点から行った．

5.1 閲覧履歴の収集とマイクロクラスタの構築の実装

実験のために，拡張機能の作成が容易である Chrome ブラウザを用いて閲覧履歴を収集した．ページ遷移やタブに関する情報の獲得には，開発者向け API（Chrome Platform APIs）と Document Object Model を利用した．ブラウザのタブ動作に関する情報 [*3] を中心に，ページへの訪問に関する情報 [*4] と関連付け，互いの情報を補完して閲覧履歴を構築した．閲覧履歴には，閲覧履歴の識別する history_id，ページへの遷移方法を表す transition，訪問を一意に識別する visit_id，閲覧したタブの識別子である tab_id，タブが属すウィンドウを識別する window_id，URL，document.referrer（referrer），遷移元の visit_id を指す referring_visit_id，新規タブを開いた際の開き元 tab_id を指す opener_tab_id，ページのソーステキストなどの情報が含まれる．

続いて，マイクロクラスタの構築についての実装を述べる．分類対象とする閲覧履歴を発生順に，$H = \{h_0, h_1, ..., h_{k-1}\}$ とする．以下の **1** から **5** の手順で，履歴 $h_i (1 \leq i < k)$ のそれぞれの連結先 h_{dst} を $h_j (0 \leq j < i)$ から探し，マイクロクラスタを構築する．はじめ，連結先がない状態を $h_{dst} \leftarrow \phi$ とする．

1. タブの再訪 h_i と tab_id, window_id, URL, visit_id[*5] が一致する h_j があれば [*6]，h_i をタブの再訪による閲覧履歴とみなし，$h_{dst} \leftarrow h_j$ とし，**5** へ．

2. ページの再訪 h_i と，tab_id, window_id, URL が一致する h_j があれば，h_i をページの再訪による閲覧履歴とみなし，$h_{dst} \leftarrow h_j$ とし，**5** へ．

3. リファラによる関連性 以下の **a** から **d** の順に，リファラによる連結候補 h_{cnd} を探す．

a. referrer の利用 h_i と tab_id, window_id が一致し，h_i の referrer と URL が一致する h_j があれば，$h_{cnd} \leftarrow h_j$ とし，**4** へ．

b. referring_visit_id の利用 h_i の referring_visit_id が指す visit_id を持つ h_j があれば，$h_{cnd} \leftarrow h_j$ とし，**4** へ．

c. opener_tab_id の利用 [*7] h_i の transition が link であり，window_id が等しく，h_i の opener_tab_id が指す tab_id を h_{i-1} が持てば [*8]，$h_{cnd} \leftarrow h_{i-1}$ とし，**4** へ．

[*2] http://mecab.sourceforge.net/

[*3] http://developer.chrome.com/extensions/tabs.html

[*4] http://developer.chrome.com/extensions/history.html

[*5] ページへの訪問履歴を一意に識別する visit_id が一致しているかどうかを確認することで，遷移を伴う閲覧であるかを判断する．

[*6] 該当する h_j が複数存在するならば，i との差が最小の j を選択．

[*7] 新規タブでページを開いた場合に referrer や referring_visit_id では対処できないことがあるため opener_tab_id を利用する．

[*8] リンクによる遷移に限定するため，直前の閲覧履歴（$j = i - 1$）

d. ドメイン一致による関連性 h_i の transition が link であり，h_{i-1} が h_i と同一の tab_id, window_id, ドメイン名を持てば，$h_{cnd} \leftarrow h_{i-1}$ とする．

4. クエリ不一致による起点の生成 h_i と h_{cnd} が検索結果ページであり，かつ，クエリが部分一致しなければ，$h_{dst} \leftarrow \phi$ とし，そうでなければ，$h_{dst} \leftarrow h_{cnd}$ とする．

5. マイクロクラスタの構築 $h_{dst} = \phi$ ならば，h_i を閲覧の起点とし，新たなマイクロクラスタとする．$h_{dst} \neq \phi$ ならば，h_i を h_{dst} と同じマイクロクラスタに属させる．

5.2 実験の詳細

研究室内の 20 代の男性 8 名と女性 1 名に，一定期間の Web 閲覧を行ってもらった．被験者を選定する際には，タブブラウザの操作に習熟し，Chrome を利用したことがあることを条件とした．Web 閲覧に関しては普段通りの閲覧を心がけてもらった．閲覧履歴を収集する期間は被験者が 1 度でも Web 閲覧した日を 1 日と数え，累積で 7 日間とした．閲覧履歴を収集した後，期間の違いが分類精度に与える影響を測るため，はじめの 1 日，はじめの 3 日，および 7 日間の 3 つの期間の閲覧履歴を被験者に渡し，「細かい意図」と「粗い意図」ごとにまとめてもらった．2 種類の粒度を明示的に指定することで，様々な粒度に合った分類が可能かを評価した．細かい意図については，閲覧履歴に共通した目的や興味があれば，それらをまとめるように指示し，粗い意図については，細かい意図のまとまり間で共通する部分があれば，それらをまとめるように指示した．意図や目的，興味についてをどう考えるか，それぞれの粒度において具体的にどの程度でまとめるかについては被験者に委ねた．意図に対する個人の考えを尊重するため，それぞれのまとめかたに関する例示は避けた．

5.3 評価観点と評価尺度

比較手法として，マイクロクラスタの構築を行わず，閲覧履歴に含まれるページ本文のみを利用して群平均法を適用したものを用いた．距離の算出などは提案手法と同様である．被験者が意図ごとにまとめた閲覧履歴と，提案手法および比較手法の最終クラスタとの一致性を Adjusted Rand Index（ARI）[8] を用いて評価する．ARI は同一対象データに対する 2 つの分類結果の類似性を測るもので，1 で完全一致，0 でランダムクラスタリングの期待値となる．

提案手法および比較手法ともに，*threshold* を 0 から 1 まで 0.01 刻みで変化させて，最終クラスタを形成し ARI を算出する．被験者の細かい意図や粗い意図に対して ARI が一定程度あれば，閾値の操作によって最終クラスタが表す閲覧意図の粒度操作ができるといえる．容易かつ自由な粒度操作を実現するためには，幅広い閾値で一定程度の ARI があり，最適解（最大 ARI）近辺が広いことが望ましい．

のみを連結対象としている．

被験者 意図	マイクロ クラスタ	閲覧履歴 番号	tabid	ページタイトル	URL（一部省略）
a	i	855	304	かわいい顔文字の総合サイト 顔文字カフェ	kaomoji-cafe.jp
a	i	856	304	シンプル短めの顔文字 顔文字カフェ	kaomoji-cafe.jp/facemark/simple/mijikame/
b	ii	857	296	(46) b3 - co-meeting	www.co-meeting.com/g/762685069635(…)
c	iii	858	308	Gmail	mail.google.com/mail/
c	iii	859	308	Gmail	mail.google.com/mail/u/0/
c	iii	860	308	Gmail	mail.google.com/mail/u/0/#inbox
b	ii	861	296	(46) b3 - co-meeting	www.co-meeting.com/g/762685069635 (…)
d	iv	862	312	purobaida - Google 検索	www.google.co.jp/search?q=purobaida+ (…)
d	v	863	312	プロバイダー - Google 検索	www.google.co.jp/search?q=purobaida+ (…)
b	ii	864	296	(46) b3 - co-meeting	www.co-meeting.com/g/762685069635(…)
d	vi	865	316	OCN光 フレッツ マンション - Google 検索	www.google.co.jp/search?q=OCN%E5%85(…)
b	ii	866	296	(46) b3 - co-meeting	www.co-meeting.com/g/762685069635 (…)
e	vii	867	320	ニュース 要約 アメリカ - Google 検索	www.google.co.jp/search?q=%E3%83%8B(…)

図 3 複数の閲覧意図が出現する閲覧履歴の例（被験者 F）
Fig. 3 Participant F's partial browsing history which contains multi browsing intentions.

閲覧の起点が意図の出現を捉えられているかに関する評価では，適合率および再現率とそれらの調和平均である F 値を用いた．適合率および再現率の評価式をそれぞれ式 (6)，式 (7) に示す．マイクロクラスタと，被験者ごとに意図に対して ARI が最大となる閾値で作成した最終クラスタ（これを，最大 ARI クラスタと呼ぶ）を用いて，閲覧の起点の評価を行った．最終クラスタにおける閲覧の起点は，各クラスタ内の最古の閲覧履歴とした．意図の出現は，被験者が意図ごとに分類した際の各まとまりにおける最古の閲覧履歴とした．

$$\text{適合率} = \frac{\text{閲覧の起点と意図の出現との一致数}}{\text{閲覧の起点の数}} \quad (6)$$

$$\text{再現率} = \frac{\text{閲覧の起点と意図の出現との一致数}}{\text{意図の数}} \quad (7)$$

マイクロクラスタでは閲覧意図を網羅的に捉えることを目指しており，閲覧の起点の再現率が高いことが期待される．被験者の意図との一致性が高い最大 ARI クラスタにおいて，閲覧の起点の再現率および適合率が一定程度あれば，意図の出現を獲得できており，閲覧意図を捉えられているといえる．

5.4 被験者による閲覧履歴の分類

各被験者の閲覧履歴の数と意図ごとのまとまりの数を**表 1** に示す．1 日目の閲覧履歴数の最小は被験者 G の 12 で，次点で被験者 I の 35 である．これらの被験者は，はじめの 1 日間では閲覧数が他の被験者と比べて少ないが，3 日間および 7 日間では他の被験者と同程度の閲覧を行っている．まとめる際の意図の捉え方や，Web 閲覧時の主な閲覧意図は被験者によって異なるため，同程度の閲覧履歴数であっても，それぞれの被験者で粗い意図の数と細かい意図の数に違いがある．被験者によっては，粗い意図の数と細かい意図の数が近いこともある．

タブを用いた複数の閲覧意図を満たすような並列的な閲覧行動と，そのときのマイクロクラスタの分類結果についての例を述べる．**図 3** は被験者 F の 3 日目における閲覧履歴の情報に加え，被験者による細かい意図の分類結果（列

表 1　被験者の閲覧履歴数と意図の数
Table 1　The number of browsing history and intentions.

期間		A	B	C	D	E	F	G	H	I	平均	標準偏差
1 日	閲覧履歴数	682	121	88	321	835	381	12	424	35	322.11	274.93
	細かい意図の数	65	10	23	26	189	25	4	19	3	40.44	55.32
	粗い意図の数	60	10	8	13	31	20	3	11	2	17.56	17.17
3 日	閲覧履歴数	1342	515	333	545	1241	1200	379	2124	761	937.78	555.13
	細かい意図の数	101	22	75	33	220	109	50	24	34	74.22	59.94
	粗い意図の数	81	20	28	17	47	98	15	14	20	37.78	29.48
7 日	閲覧履歴数	2589	1044	1232	1262	3522	1935	1676	4285	1681	2136.22	1052.88
	細かい意図の数	239	36	299	55	353	228	225	27	79	171.22	116.04
	粗い意図の数	218	30	55	23	106	204	36	17	36	80.56	74.02

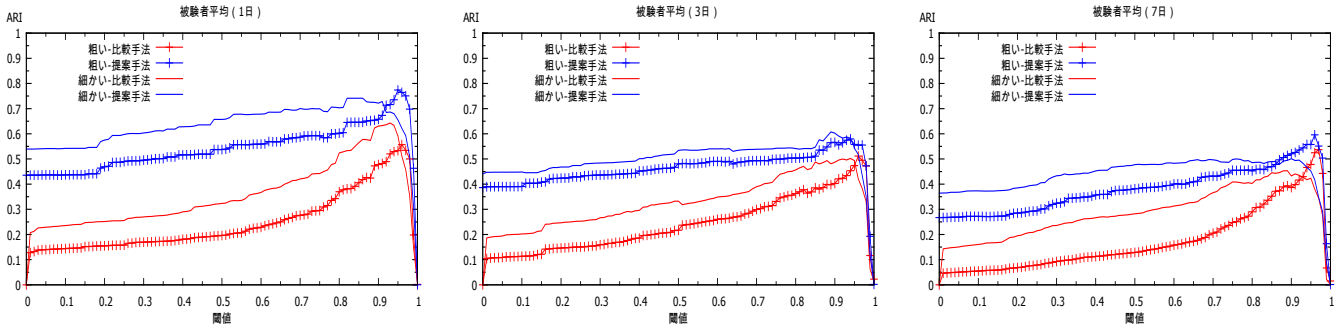

図 4　提案手法と比較手法の閾値の変化に伴う平均 ARI
Fig. 4　A comparison of averaged ARI between proposed and comparative method.

名「被験者意図」）とマイクロクラスタによる分類結果を記している．ここでは，被験者が同一の意図とした閲覧履歴を同一の記号（a〜e）で表し，同一のマイクロクラスタに属す閲覧履歴は同一の記号（i〜vii）で表す．閲覧履歴は時系列昇順に並べている．閲覧意図が頻繁に変化する時，例えば，閲覧時刻の近さの指標を用いて分類すると，番号が 860 と 861，862 の閲覧履歴が並んでおり，意図 c，b，d は混在しえる．一方，マイクロクラスタでは，閲覧履歴の関連性を用いることで被験者意図に沿ったかたちでの分類ができている．意図 d のみ，3 つのマイクロクラスタに分かれているが，マイクロクラスタは閲覧意図を最細粒度で捉えることを目的としているためである．

5.5　一致性の観点からの評価結果と考察

閾値の変化に伴う ARI の全体的な傾向を見るため，被験者の平均 ARI を示す．**図 4** は，1 日間，3 日間，7 日間の各期間における，全被験者の ARI の平均を表す．平均 ARI は，3 つの期間のほとんどの閾値で比較手法より提案手法が高い結果となった．提案手法，比較手法ともに長期間になるにつれて全体的に ARI が低くなる傾向がある．一因として，被験者が長期間の大量の閲覧履歴を閲覧時の意図ごとに正確にまとめることが困難だったことが考えられる．提案手法において，$threshold = 0$，すなわちマイクロクラスタの時点で ARI が一定程度あり，閾値が上がるにつれて ARI が向上している．特に粗い意図に対して，比較手法は 0.95 程度の狭い範囲の閾値でのみ ARI が高まる傾向がある一方，提案手法ではマイクロクラスタの構築によって，より低い閾値から幅広い閾値で ARI が高い．比較手法は ARI が高い閾値の範囲が狭く，閾値の操作によって最適解の選択が困難であると予想される一方，提案手法では最大 ARI 付近の閾値が広く，閾値を操作によって最適解に到達しやすいと考えられる．

被験者ごとの最大 ARI の平均について，提案手法が比較手法に対して有意に高いかについて，有意水準を 5 %として対応のある t 検定を行った．1 日間の粗い意図と細かい意図，および 7 日間の粗い意図における p 値はそれぞれ，0.020，0.049，0.024 となり，有意差があった．一方，3 日間の粗い意図と細かい意図，および 7 日間の細かい意図における p 値はそれぞれ，0.087，0.388，0.270 となり，有意差がなかった．特に，細かい意図に対して p 値が高い傾向にある理由は，提案手法で ARI が向上しなかった被験者 2 名にあると考えられる．これらの被験者については，後述する，他にも，細かい意図や長期間になるに連れて，p 値が高まる一因として，前述のとおり，被験者が意図ごとにまとめた際の正確性が下がり，マイクロクラスタで捕捉を目指した閲覧意図と一部が異なったために，提案手法の一致性が低下したことが考えられる．

比較手法では，ページ本文を取得できなかったために分類できなかったものが，7 日間の閲覧履歴において平均で 150 程度あった一方，提案手法では平均で 12 程度であっ

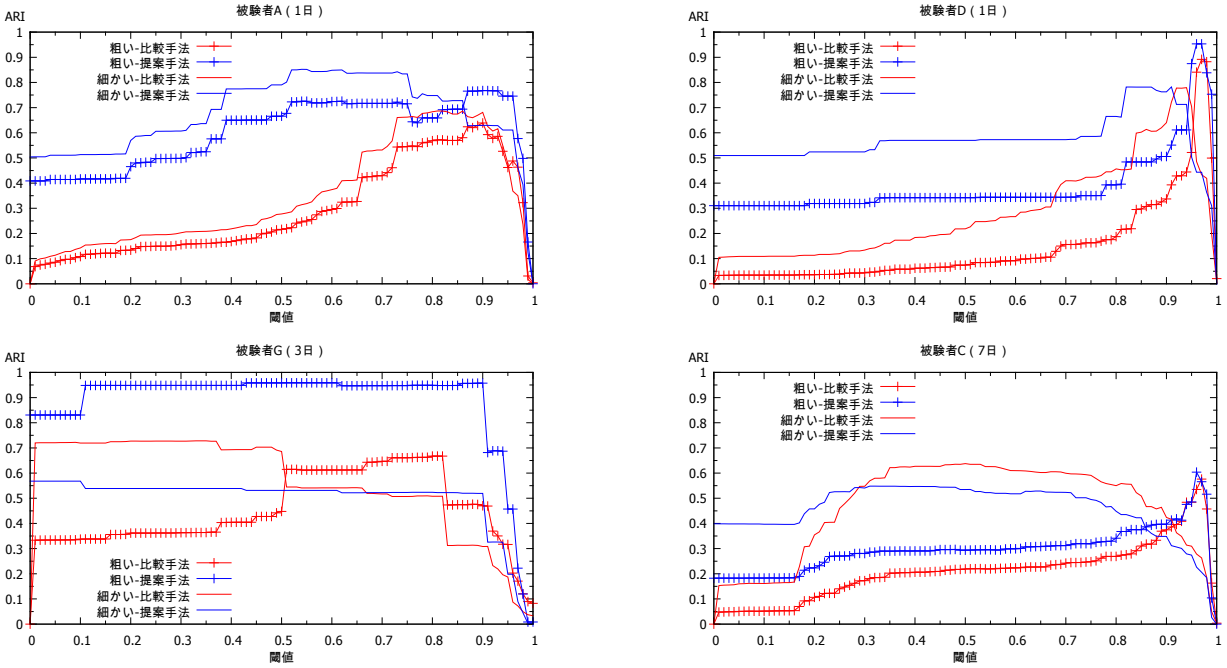

図 5　被験者 A, D, G, C の閾値の変化に伴う ARI

Fig. 5　A comparison of averaged ARI with browsing history of A, D, G and C.

表 2　閲覧の起点に関する評価

Table 2　Evaluation of the starting points of browsing.

		マイクロクラスタ			最大 ARI クラスタ（提案手法）			最大 ARI クラスタ（比較手法）		
期間	粒度	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
1 日	粗い	0.402	0.888	0.525	0.696	0.777	0.719	0.387	0.672	0.467
	細かい	0.546	0.752	0.600	0.736	0.699	0.704	0.478	0.639	0.520
3 日	粗い	0.282	0.796	0.392	0.516	0.613	0.546	0.211	0.569	0.295
	細かい	0.429	0.670	0.472	0.657	0.534	0.579	0.300	0.602	0.381
7 日	粗い	0.206	0.774	0.297	0.459	0.547	0.471	0.161	0.445	0.223
	細かい	0.361	0.687	0.426	0.550	0.511	0.512	0.251	0.519	0.320

た．マイクロクラスタの構築によって本文が取得できないページの分類が可能になった．

続いて，各被験者の考察を行う．被験者 A，D，G，C の閾値と ARI の関係を**図 5** に示す．被験者 A および D の 1 日間の結果では，提案手法において，マイクロクラスタで ARI が高く，閾値を上げていくと最大 ARI も高くなる．比較手法では，狭い閾値の範囲でのみ ARI が高いが，提案手法の ARI はより低い閾値から高い．被験者 D は閲覧意図をページ内容によって区別することが容易であったため，比較手法および提案手法ともに最高 ARI が高い．

一方，提案手法で ARI が向上しなかった被験者が 2 名いた．被験者 G の 3 日間は，細かい意図においてマイクロクラスタで最大 ARI となった例である．この被験者の 3 日間のマイクロクラスタの数は 32 であったが，細かい意図の数は 50 とそれよりも多かった．被験者 G は，検索結果ページからあるページに遷移し，検索結果ページへ再訪したのち，さらに別のページを閲覧した際に，はじめの検索結果ページと遷移先ページと，検索結果ページの再訪と別ページを異なるまとまりとしていた．被験者の最細粒度よりもマイクロクラスタが粗くなったため，階層的併合により，細かい意図との ARI が下がる．被験者 C の 1 日間の細かい意図でも同様に，マイクロクラスタで最大 ARI となった．

図 5 において，被験者 C の 7 日間の細かい意図で，提案手法の ARI が比較手法と比べて低い．この被験者は，1 週間で細かい意図が 299 あったが，マイクロクラスタの数はそれよりも少ない 194 であった．これは，被験者が考える意図が，我々が分類対象とする閲覧意図と異なっていたためである．被験者 C は，「開いているページを確認するためにタブを開いた」「目的のタブと誤って他のタブを開いた」「閉じるためにタブを開いた」など，短時間のタブ操作を意図の 1 つとして捉えていた．タブ操作に関する意図ごとにまとめられた場合，提案手法の ARI が低下する．このような意図は，今回我々が事前に想定した閲覧意図の

表 3　マイクロクラスタを用いた閲覧の起点に関する被験者群の比較
Table 3　Comparison between the group of partcipants using micro-clusters.

	意図の数が多い被験者群			意図の数が少ない被験者群		
期間	適合率	再現率	F 値	適合率	再現率	F 値
1 日	0.784	0.571	0.640	0.478	0.804	0.588
3 日	0.713	0.427	0.533	0.348	0.740	0.454
7 日	0.541	0.421	0.470	0.310	0.763	0.414

範囲外であったが，今後は例えばごく短期のタブの閲覧は閲覧履歴から省くなどの処理が必要である．

この 4 人の例から，被験者ごとに最適な閾値が異なることがわかる．2 章で述べたように，ユーザに対して適切な粒度を機械的に捉えることは困難であり，閾値を機械的または経験的に決定する手法を適用すべきでない．

5.6　閲覧の起点に関する評価結果と考察

マイクロクラスタ（第 1 段階のみの分類結果）および最大 ARI クラスタと，被験者が意図ごとにまとめた結果を用いて閲覧の起点を評価した．評価結果の全被験者の平均を**表 2** に示す．マイクロクラスタでは再現率が一定程度あり，閲覧の起点で意図の出現を網羅的に捉えられている．一方，マイクロクラスタは閲覧意図を最細粒度で捉える目的で構築したため，その適合率は再現率に比べて低い．提案手法の最大 ARI クラスタの結果を見ると，マイクロクラスタから再現率の低下を抑えつつ，適合率が向上している．階層的併合によって，マイクロクラスタで網羅的に獲得した意図の出現を正確に捉えられるようになった．

閲覧意図とその推移を捉えるためには，一致性が高く，かつ意図の出現を網羅的かつ正確に獲得できる必要がある．提案手法において，被験者の意図との一致性が最も高まる最大 ARI クラスタにおいて，閲覧の起点の再現率および適合率は，比較手法に比べて高い．このことから，提案手法の分類は，閲覧意図を捉えられているといえる．

意図の出現を捉えるためには，マイクロクラスタで閲覧意図を最細粒度で捉えることが重要である．そこで，細かい意図の数がマイクロクラスタの数より多かった被験者 2 名（被験者 C と G）と，細かい意図の数がマイクロクラスタの数よりも少なかった被験者 7 名を区別して閲覧の起点の評価を行った．結果を**表 3** に示す．マイクロクラスタよりも意図を細かく捉えた被験者群では，そうでない被験者群と比べ，再現率が低い．これは．5.5 節で述べたように，本論文で分類対象とした閲覧意図と異なる考え方をしていたためであり，意図の出現をさらに網羅的に捉えるには，今後，このようなユーザに対応する必要がある．

4.2.1 節で述べたクエリ不一致による起点の生成を 1 週間の閲覧履歴を用いて評価したところ，生成しない場合と比べて再現率が全被験者平均で 0.05 程度向上し，最大で被験者 D において 0.2 向上した．この操作によって，検索時の意図の出現を捉えられるようになった．

6.　おわりに

本論文では，新たな意図の出現を捉えることで，ユーザ自身が考える閲覧意図を反映した閲覧履歴の分類手法を提案した．提案手法では，閲覧履歴を 2 段階で分類する．第 1 段階では，閲覧履歴間の関連性によってユーザの閲覧意図を最細粒度で表すと期待されるマイクロクラスタの構築を行い，第 2 段階では，マイクロクラスタに含まれる本文を利用し，群平均法を適用することでマイクロクラスタの階層的併合を行った．これによって閲覧意図の階層構造を捉え，最終的な分類結果が表す閲覧意図の粒度操作を容易かつ自由に行える．

閲覧履歴を被験者自身が意図ごとに 2 種類の粒度でまとめたものと提案手法の分類結果の一致性を ARI によって評価したところ，ページ本文のみによる比較手法よりも幅広い閾値で一致性が高いことを示した．さらに，マイクロクラスタで意図の出現を一定程度網羅的に捉えることができ，マイクロクラスタの階層的併合によって正確性が高まることを確認した．

今後の課題は，閲覧意図をより細かく捉えているユーザに対応して，閲覧時間やページ遷移の広がりなどを考慮し，マイクロクラスタにおける閲覧の起点の網羅性を高めることや，目的のタブを開く行為やタブを閉じるために開く行為などを判定し，タブに関する操作を捉えることである．

参考文献

[1] 長野翔一，市川裕介，小林　透：短期的な興味プロファイル構築に向けたウェブ閲覧履歴のクラスタリング方式の提案，電子情報通信学会論文誌，Vol. 95, No. 4, pp. 734–746 (2012).
[2] 飯野亜耶，奥野　拓：履歴分類の提示とアノテーションによるリファインディング支援，インタラクション，Vol. 2012, No. 3, pp. 545–550 (2012).
[3] Capra, R., Pinney, M., Manuel, A. and Perez-Quinones: Refinding is Not Finding Again, Technical report, Virginia Tech (2005).
[4] Montgomery, A. L., Li, S., Srinivasan, K. and Liechty, J. C.: Modeling Online Browsing and Path Analysis Using Clickstream Data, *Marketing Science*, Vol. 23, pp. 579–595 (2004).
[5] 長野翔一，高橋寛幸，中川哲也：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，情報処理学会研究報告，Vol. 2008, No. 90, pp. 65–70 (2008).
[6] 丸山　修，大島宗哲，鍾　寧：特異性指向によるクリックストリームマイニングに関する研究，電子情報通信学会技術研究報告，Vol. 108, No. 384, pp. 85–90 (2009).
[7] 吉田拓磨，中村明順，安積卓也，西尾信彦：閲覧の起点と関連性に注目したユーザ意図の推移追跡，第 1 回 WI2 研究会，Vol. 1, pp. 51–52 (2013).
[8] Hubert, L. and Arabie, P.: Comparing partitions, *Journal of classification*, Vol. 2, No. 1, pp. 193–218 (1985).