1F2-8

# コンテキストを変化させる閲覧履歴の抽出

An approach for detecting the change of intention from browsing-behavior

長野 翔一　中川 哲也　高橋 寛幸
Shouichi Nagano　Tetsuya Nakagawa　Hiroyuki Takahashi

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

We propose an approach for detecting the change of intention from user's browsing behavior. It is necessary to treat the user's intention accurately in information explosion. However, treating dynamic intention is difficult for a conventional method, as behavior targeting model. We hypothesize that "The factor in which the user's intention is changed in the access-log exists", and defined this as the context-driven-history. For detecting user's intention change in browsing-history, we analyze each of browsing-history based on the similarities, and show the importance of extracting context-driven-history in this paper, and we report on result of an experiment to support the hypothesis.

## 1. はじめに

近年の情報爆発時代における情報提供サービスは検索分野におけるパーソナル化，状況に応じた検索 (画像検索，位置情報検索，ニュース検索など) の増加，広告分野における行動ターゲティングのシェア増加 [e-marketer 07]，など，ユーザのニーズの多様化に伴い個人特化・状況特化の傾向が強くなっている．

本稿では，情報の個人化を進めるため，ユーザのダイナミックな要求の変化に対応した情報提供方式を目指し，ユーザの要求把握に関する課題に取り組む．

そのために，新たな要求把握方法として，要求の変化が閲覧履歴上のコンテキストの変化として表出することを前提に，その変化の原因となる要素の抽出方式を提案する．さらに，有効性の検証についても実験結果を報告する．

## 2. 背景

情報提供方式が行なう処理には大きく分けて次のような二つのステップがある．

**Step 1** ユーザの要求を捉える．

**Step 2** ユーザの要求と適切な情報をマッチングする．

Step 2 の要求と情報をマッチングする技術についてはエンジンの性能向上に伴って高い精度が確保されるようになった．Step 1 のユーザの要求把握方法は個人特化・状況特化の情報提供方式において重要な核となっており，本稿が取り組む技術である．

従来の要求把握方法で最も有名なクエリ入力型は，入力や前提知識を必要とし，ユーザへの負担が大きいとされている．この課題を解決するためにコンテンツターゲティング型，行動ターゲティング型といったユーザの入力を必要としない暗示的手法が生み出されてきた [土方 04]．コンテンツターゲティング型は閲覧中のウェブページの内容をユーザの要求として取得する手法で，閲覧中のウェブページが必ずしもユーザの要求と一致していないという問題があった．一方，行動ターゲティング型は履歴から興味プロファイルを構築し，ユーザの要求を把握する手法で，コンテンツターゲティング型と比較すると要求把握精度が高いという長所を有している．しかし，ユーザの要求が変化し，普段探したことのない情報を探すとき，行動ターゲティング型は要求を捉えることができないという問題がある．本稿では，行動ターゲティング型の長所を保ちながら，ユーザの要求変化に対応できる要求把握方法を提案する．

連絡先: 長野翔一，日本電信電話株式会社，
〒 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

## 3. 要求変化を促進させる履歴抽出の提案

### 3.1 提案内容

ダイナミックな要求変化を捉えるために，本稿では行動ターゲティング同様，閲覧履歴を活用する．そして，特定の行動がユーザの要求を変化させているという仮説を設定し，閲覧履歴中からユーザの要求を変化させる行動を抽出する．

**研究仮説** 要求を変化させる特定の行動が存在する．

仮説における『特定の行動』の履歴を変化促進履歴と定義し，コンテキストを変化させる履歴として抽出に取り組む．

従来の要求把握技術は要求の変化には着目せず，要求の結果 (行動) から要求を推定していたが，本技術は閲覧履歴中から要求の変化を引き起こす原因となる要素を特定することでユーザの要求を把握しようと試みるものである．

また，行動ターゲティングには「ユーザの行動は単一の要求から生成されており，要求は一定期間持続的・集中的に存在する」という前提がある．複数の要求から行動が生成される場合，ユーザの行動モデルが複雑化するため，今回の取り組みではこの前提を自明のこととし，採用する．この前提を採用するとき，ウェブ閲覧行動においてユーザの要求は連続して取得される類似閲覧履歴群に反映される．変化促進履歴は類似閲覧履歴群を前後に有しているため，前後双方の要求の特性 (意味的類似性) を有しており，コンテキストの変化を反映している．そのため，要求の変化の変遷を捉えることが可能となる．

本稿では変化促進履歴を抽出し，ユーザの閲覧行動におけるコンテキストの変化を捉えることで，要求の変化を把握する手法を提案する．

### 3.2 技術の説明

要求変化に対応する技術として興味プロファイルの重み付け技術がある．要求変化への対応，ノイズの影響という観点で提案手法と興味プロファイルの重み付け技術を比較する．興味プ

ロファイルの重み付け技術は，新しいコンテンツを古いコンテンツより重視することで要求の変化に対応している．しかし，新しいコンテンツの重みを増やすほど，構築したプロファイルは新しく出現したノイズの影響を大きく受けることとなる．提案方式はコンテキストの変化に着目しており，変化促進履歴で変化を把握し，変化促進履歴以降の最も新しい類似閲覧履歴群を対象とした，要求に関するプロファイルを構築することでノイズの影響を抑えることができる．即ち，要求変化の原因となる閲覧履歴を抽出することで，閲覧履歴を要求が変化する毎に分割して扱うことが可能となる．

## 4. 仮説検証

### 4.1 検証事項

研究の仮説を検証するために，変化促進履歴の存在と抽出可能性を裏付ける被験者実験を行なった．

### 4.2 検証実験

**実験概要**

ウェブリテラシーを有した 24〜26 才の被験者 5 名による実験を行なった．被験者は Wikipedia サイト内を閲覧履歴 (日時，タイトル，URL) を取りながら約 1.5 時間巡回し，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．マーキングされた閲覧履歴が変化促進履歴であるかどうかを検証する．

**分析方法**

1. 被験者のマーキングを利用して同一の要求から生成された連続的な閲覧履歴群をまとめる．この閲覧履歴群を要求クラスタと定義する [*1].
2. 「マーキングされた閲覧履歴のウェブページ」と「マーキング直後の要求クラスタ」の類似度 $p$ を算出する．
3. 「マーキング閲覧履歴を含む要求クラスタ」と「マーキング直後の要求クラスタ」の平均類似度 $q$ を算出する．
4. $p > q$ を満たすものを抽出可能な変化促進履歴とする．
5. 被験者ごとに要求変化を起こした閲覧履歴 (マーキングされた履歴) 中に占める変化促進履歴の割合を算出する．

### 4.3 実験結果の分析

- マーキングされた履歴中に占める変化促進履歴の割合は 73.3%．つまり，半数以上の要求変化が抽出可能な変化促進履歴によって引き起こされており，変化促進履歴が存在する．
- 変化促進履歴による要求変化の起こりやすさは要求変化の頻度には大きく影響されない．
- ユーザは一時間に平均 4.1 回要求を変化させており，変化促進履歴は一時間に 3.1 回程度の頻度で起こる．

以上のように，実験の結果，研究仮説が支持された．

### 4.4 事例紹介

図 1 は被験者 A の閲覧履歴の一部である．以降，図を使用して具体的に説明する．被験者 A の No.31–34 における要求は奈良地域に関するウェブページであり，No.35–42 における要求は電磁気学に関するウェブページである．被験者 A は No.34 を要求が変化したポイントとしてマーキングしている．仮説に従えば No.34 は奈良地域に関する要求から生成されている

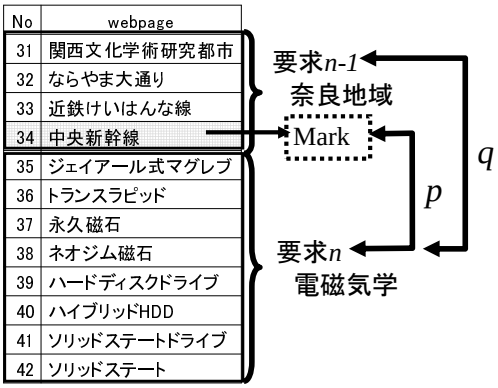


図 1: 被験者 A の閲覧履歴の一部

が，直後の電磁気学に関する要求の生成物 (No.35–42) の持つ特性 (文書の類似度) を自身の要求クラスタ内 (奈良地域に関する要求) の他の閲覧履歴に比べて強く有していると考えられる．つまり，$p = sim($ テキスト No.34, テキスト No.35–42$)$, $q = sim($ テキスト No.31–34, テキスト No.35–42$)$ と定義される $p, q$ の値が $p > q$ を満たすとき，ユーザがマーキングした No.34 は奈良地域に関する要求を電磁気学に関する要求へ変化させた変化促進履歴であるといえる．ここで，ベクトル空間法 [中川 03]， [Salton 83] により，意味的類似度を算出すると $p$, $q$ はそれぞれ $p = 0.092$, $q = 0.043$ であった．ゆえに，$p > q$ が成立し，ユーザがマーキングした要求の変化点『No.34 中央新幹線』が変化促進履歴と判定される．

このように，自身が属する要求 $n-1$ に影響し，要求 $n-1$ を要求 $n$ に変化させる性質を持つ閲覧履歴を変化促進履歴として抽出した．

## 5. まとめと今後

本稿ではウェブページの閲覧履歴を対象に変化促進履歴という要求の変化の原因となる要素に着目し，変化促進履歴の抽出技術を提案した．さらに，その提案に基づき仮説となる変化促進履歴の存在と抽出可能性を仮説検証実験を通じて明らかにした．今後は提案技術を応用し，ユーザのマーキングのない閲覧履歴からの変化促進履歴抽出，タブブラウザに代表される要求が並在する閲覧行動への適用拡大，等を検討する．

## 参考文献

[e-marketer 07] e-Marketer, Behavioral Advertising on Target... to Explode On http://www.emarketer.com/ (2007).

[土方 04] 土方嘉徳: 情報推薦・情報フィルタリングのためのユーザプロファイリング技術, 人工知能学会論文誌, Vol.19, No.3 pp. 365-372 (2004).

[中川 03] 中川裕志, 湯本紘彰, 森辰則: 出現頻度と連接頻度に基づく専門語抽出, 自然言語処理, Vol.10, No.1 pp. 27-45 (2003).

[Salton 83] Salton, G. and McGill, M. J.: Introduction to Modern Information Retrieval, McGraw-Hill (1983)

*1 $n$ 番目の要求クラスタは，$n-1$ 番目にマーキングされた閲覧履歴の次の閲覧履歴から始まり，$n$ 番目にマーキングされた閲覧履歴で終わる．