5S-4

SOMを用いた単語関連性に基づくユーザの嗜好抽出手法の提案

河野 慎[†] 蛭田 慎也[‡] 伊藤 友隆[‡] 米澤 拓郎[‡] 中澤 仁[†] 徳田 英幸[†]
慶應義塾大学 環境情報学部[†] 慶應義塾大学大学院 政策・メディア研究科[‡]

1. はじめに

近年 Amazon を始めとした E コマースサイトにおいて，ユーザの嗜好を用いた推薦システムが普及している [1]．推薦アルゴリズムの一つに協調フィルタリングがあり，これは推薦対象のユーザと類似した評価をしているユーザを検索し，検索したユーザの評価情報に基づいて推薦する [2]．この協調フィルタリングを Web サービスで利用する場合，主に商品の情報やユーザの閲覧・購入履歴情報を特徴量として用いる．このような閲覧履歴を用いることで，ユーザの嗜好を抽出することが可能である．

従来の推薦システムでは，ユーザが購入した商品や閲覧履歴などから推薦を行なっている．しかし，これらの情報はそれぞれのサービスで収集されたものであるため，別サービスとは共有されない．そのため，ユーザが訪れることの少ないサイトなどでは，嗜好を正しく反映していない推薦が行われることがある．そこで本研究では，ユーザが日常的に行なっている Web ページの閲覧履歴からユーザの嗜好を抽出する．さらに，抽出した嗜好をもとにして，同一の Web サービス内の履歴に偏らない推薦システムを構築し，サービスや情報の横断的な推薦を可能にすることを目指す．

本研究の提案手法を利用した推薦の例として，飲食店レビューサイトと料理レシピ提供サイトを組み合わせた場合がある，ユーザが普段飲食店レビューサイトで検索したり，実際に訪れたりしている店舗の情報を取得する．そして料理レシピ提供サイトでレシピ検索を行う際に，取得した情報に基づいて推薦を行うことが可能になる．

2. 関連研究

Web の閲覧履歴を用いた既存研究として，高須賀ら [3] は URL を元に類似ユーザを抽出し，これを基に推薦を行っている．この手法は協調フィルタリングを用いており，類似ユーザに基づいているため，この類似ユーザによって推薦が変わってしまう問題がある．本研究では，推薦対象のユーザの情報のみを利用し，この問題を解決する．

また大森ら [4] は URL からページの本文を抽出し，それらの中からマッチングキーワードを利用した推薦を行なっている．しかし，この手法では偶然見てすぐに別のページに移動したページや，逆に長時間見ていたページも同等の重みで解析に使用しているため，閲覧ページに対してユーザがどの程度興味を持っていたかが考慮されていないという問題がある．本研究では，Web ページの閲覧履歴だけでなく，閲覧していた時間やタイミングを考慮したユーザの嗜好抽出手法を提案する．

Preference Extraction of User Based on the Word Association Using SOM

[†]Makoto Kawano, Jin Nakazawa, Hideyuki Tokuda
[†]Faculty of Environment and Information Studies, Keio University
[‡]Shinya Hiruta, Tomotaka Ito, Takuro Yonezawa
[‡]Graduate school of Media and Governance, Keio University

3. 嗜好抽出アルゴリズム

ユーザの Web ページ閲覧履歴を取得するため，本研究では Google Chrome Extension を実装する．この Extension は，ユーザが Google Chrome である URL にアクセスすると，サーバにその URL とその時刻が送信する機能をもつ．また Google Chrome でタブを閉じたり，別のアプリケーションがアクティブになったりした際にも，そのイベントとその時刻をサーバに送信する．閲覧時間は式 (1) によって求まる．

$$t_g = t_{p_{i+1}} - t_{p_i} \tag{1}$$

ここでユーザ d が閲覧した Web ページの集合を $P^d = (p_1, p_2, p_3, ..., p_n)$ とする．t_g は閲覧時間，t_{p_i} はページ p_i の閲覧開始時刻を表す．取得した URL から HTML をパース・形態素解析を行う．本研究では MeCab を使用し，名詞を抽出する．そして抽出した単語と閲覧時間を用いて単語の特徴ベクトルを抽出する．特徴ベクトルを式 (2) に示す．

$$\vec{w_i} = (appear, tf/idf, time, hour) \tag{2}$$

ここで $\vec{w_i}$ は単語 i の特徴ベクトル，$appear$ は単語の出現回数，tf/idf は単語の tf/idf 値，$time$ は閲覧時間，$hour$ は閲覧された時刻を表す．そしてこの特徴ベクトルを入力値にして SOM(Self Organization Map) を用いて，単語を 2 次元にマッピングを行う．マッピングされた単語は特徴ベクトルが類似しているほど近くに配置され，密に集まっている単語がそのユーザの嗜好を表現しているといえる．

4. 実験

4.1 実験手順

本研究で提案する嗜好抽出手法を評価するため，実際に被験者に利用してもらい，被験者の嗜好抽出を試みた．被験者は 6 名で全員 20 代の男性で，実験期間は 2012 年 12 月 27 日から 2013 年 1 月 7 日までとした．被験者に Google Chrome Extention のインストールと，サーバに CGI とデータベースを設置してもらう．その状態で普段通り Web 閲覧を行ってもらい，実験期間終了後にデータベースを提供してもらう．収集した情報から単語の特徴ベクトルを抽出し，これを入力値にして SOM_PAK を用いる．マッピングされた図をそれぞれのユーザに見てもらい，アンケートに答えてもらう．アンケートはで，(1) 実験期間中どのような Web サイトを閲覧したか (2) 図にはそれが反映されているかという設問をすべて自由記述形式で設定した．

4.2 実験結果

図 1,2 に実験結果のうち特徴的なものを 2 つ，図 3 に図 2(b) 中の枠部分の拡大※を示す．それぞれの図の (a) の図は単語の特徴ベクトルに閲覧時間を入れておらず，(b) の図では閲覧時間を入れている．なお，図 1,2 はそれぞれ別の実験協力者の実験結果であり，それぞれ被験者 A，被験者 B とする．

また実験後のアンケートでは，実験期間中，被験者 A は”特に調べ物などはせずに Yahoo ニュースを端から順に満遍なく閲覧”と，被験者 B は”まとめサイトなどでサッカー関連のサイトを閲覧”と答えている．

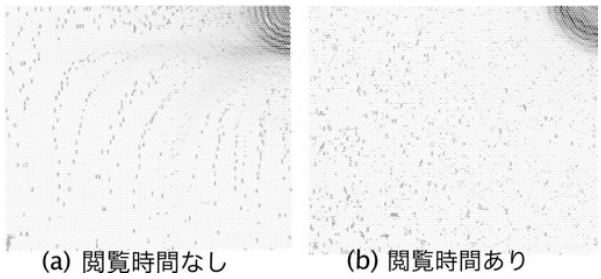

図 1: 被験者 A の嗜好抽出結果

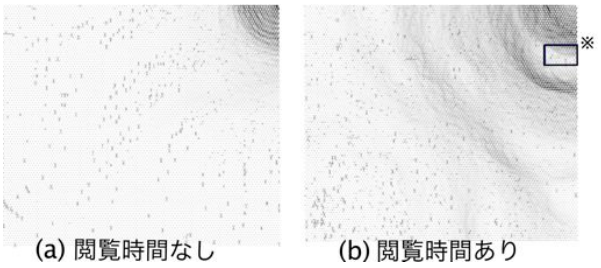

図 2: 被験者 B の嗜好抽出結果

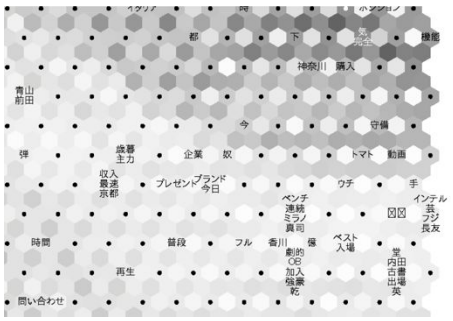

図 3: 被験者 B/閲覧時間考慮結果の一部

4.3 考察

それぞれの被験者について嗜好抽出結果とアンケートから考察を行う．被験者 A については，特に好きなものを Web で見たりしていないため，本来満遍なく単語が配置されなければならないが，図 1(a) では随所に集まって配置されているのが散見される．それに対し，図 1(b) では単語が満遍なく配置されており，実験期間中における被験者 A の嗜好を抽出できている．

被験者 B についてはサッカーに関する単語などが集まって配置されていなければならない．しかし図 2(a) では満遍なく配置されてしまっている．それに対し，図 2(b) では単語が濃淡による境界で分類されており，実際に拡大して (図 2(b) 中※，図 3) 見てみるとサッカーに関連した単語が集まってきており，被験者 B の嗜好を抽出できている．

その他の被験者についてもマッピングされた図とアンケートを比較した結果，被験者 B と同様に実験期間中意識的に閲覧していたページの単語が集まって配置されているのが見受けられた．

以上から，Web ページ閲覧時間を考慮に入れた本研究の提案手法は有効であるといえる．しかし図の全体を通して見ると，あまり特徴的ではない単語などが多く存在する．これは URL から HTML をパースをしてくる際に，本文だけでなくフォームの文字をパースしたりしてしまっているのが原因と考えられ，解決するには HTML パーサの改良が必要である．

また今回の結果を得るにあたって実験期間が 10 日間ほどであったため，長期的に利用した場合にどうなるかを試す必要がある．その際，今回の結果とは逆に閲覧時間を考慮しないほうが嗜好抽出できる可能性もあり，これの確認は今後の課題である．

5. まとめ

推薦システムの中でも協調フィルタリングは有効であり，多くの EC サイトなどで使われている，しかし推薦に利用される情報は同一の Web サービス内で閉じてしまっており，複数のサービスで横断的に推薦できない．そこで本研究では Web サービスで横断的な推薦を可能にするため，嗜好を抽出する手法を提案した．Web ページの閲覧履歴と閲覧時間を取得し，SOM を用いてその単語の関連性からユーザの嗜好の抽出を試みた．その結果，閲覧時間を考慮することで，ユーザの嗜好を定性的に抽出できたが，定量的な評価と抽出した嗜好を基に推薦を行うことが今後の展望として挙げられる．

参考文献

[1] 小野智弘，麻生英樹，本村陽一：情報・コンテンツのレコメンド技術と課題，電子情報通信学会誌，Vol.94，No.4，pp.310-315(2011)

[2] 大杉直樹，門田暁人，森崎修司，松本健一：協調フィルタリングに基づくソフトフェア機能推薦システム，情報処理学会論文誌，Vol.45，No.1，pp267-278(2004)

[3] 高須賀清隆，丸山一貴，寺田実：閲覧履歴を利用した協調フィルタリングによる Web ページ推薦とその評価，情報処理学会研究報告．データベース・システム研究会報告，2007，No.65，pp.115-120(2007)

[4] 大森慎吾，宇野達也，大野成義：閲覧履歴を利用した Web ページ推薦の SOM による視覚化，データ工学と情報マネジメントに関するフォーラム，2010