

閲覧履歴を利用した Web ページ推薦の SOM による視覚化

大森 慎吾[†] 宇野 達也[‡] 大野 成義[‡]

[†] [‡] 職業能力開発総合大学校情報システム工学科 〒229-1137 神奈川県相模原市橋本台 4-1-1

E-mail: [†] h8ptu@yahoo.co.jp, [‡] {uno, ohno}@uitech.ac.jp

あらまし 閲覧履歴からキーワードを抽出し、そのキーワードに関係するページ群を SOM によって視覚的に提示することで Web ページを推薦するシステムを提案する。閲覧履歴はユーザの興味を知る有効な情報源であり、これを用いた Web ページ推薦のシステムは色々と提案されている。本研究ではマイニングする手法として SOM を用い、その結果を 2 次元平面に配置することで閲覧履歴と推薦ページの関連性をも提示する。プロトタイプシステムによって実験を行い、閲覧履歴に関連性のあるページが推薦されていることを確認した。

キーワード 閲覧履歴, SOM, Web ページ推薦, 視覚化

Web Page Recommendation using Browsing History by SOM Visualization

Shingo OHMORI[†], Tatsuya UNO[‡], and Shigeyoshi OHNO[‡]

[†] [‡] Department of Information System Engineering, Polytechnic University

4-1-1, Hashimotodai, Sagami-hara-shi, Kanagawa 229-1196 Japan

E-mail: [†] h8ptu@yahoo.co.jp, [‡] {uno, ohno}@uitech.ac.jp

Abstract This paper attempts to propose the Web page recommendation system presenting related pages which be visualized by SOM based on the key word extracting from the browsing history. The browsing history is an effective source to pick up users' interest and have been used by many Web page recommendation systems. This study chooses SOM as a technique for data mining, and plots the result in two dimension plane presenting relativity in the browsing history and recommended pages. According to experiments on the prototype system, it was confirmed that the recommended pages are related to the browsing history.

Keyword Browsing History, SOM, Web Page Recommendation, Visualization

1. はじめに

1.1. 背景

インターネットが普及した現在、そこに存在する Web ページ数は膨大なものとなり、我々は意図する情報を得る過程で様々な Web ページを閲覧している。その閲覧履歴はユーザの利用するブラウザによってコンピュータ内にデータとして保存されている。インターネットを利用する多くのユーザは何らの意図や目的を持って Web ページの閲覧を行っており、個人の閲覧履歴はユーザの興味を知る上で有効な情報源となりえる。このようなユーザの嗜好や目的に基づく推薦システムの例として、商品の購買履歴を基にしたオンラインショッピングの商品推薦が挙げられる。また購買履歴の代わりに Web ページの閲覧履歴を利用した Web ページ推薦システムが開発されてきたが、履歴の扱い方や推薦情報の取得手法は多種多様であるため、現在も多くの推薦手法に関連する研究[1][2]が進められている。

しかし、これまでの提示手法は主に推薦対象のリスト化であり、個人の閲覧履歴と推薦される Web ページの関連性が明示的に提示されているわけではない。

1.2. 研究目的

本研究では推薦ページ群と履歴ページを SOM によって視覚的に提示することで Web ページを推薦するシステムを提案する。推薦ページはユーザが利用しているブラウザの閲覧履歴から特徴となるキーワードを抽出し、そのキーワードに関係する Web ページを用いる。そして SOM を用いることで 2 次元平面に配置された履歴ページと推薦ページの関連性を明示し、履歴ページと関連を持つ Web ページを推薦候補としてユーザに提示することを目的とする。

これにより閲覧履歴の傾向確認、更に履歴に関連する Web ページの発見をユーザに提供できると考えられる。

2. 関連研究

2.1. 他の推薦システムとの比較

従来の推薦システムは大きく2つに分類[1]される。1つはコンテンツベースのフィルタリングを用いたシステムであり、もう1つは協調フィルタリングを用いたシステムである。本研究では、ユーザ個人の閲覧履歴のみを用い、協調のための複数のユーザ履歴を必要としないこと。また、閲覧ページの記述内容を解析することから、協調フィルタリングでなくコンテンツベースのフィルタリングシステムに分類できる。

閲覧履歴ページを解析して得られたユーザの嗜好を、数値化してプロファイリングするとき、ベクトル空間モデルを使う。特徴量を空間軸とし多次元空間上に表現する。本研究では、人の好みは好き、嫌いといった単純なものではなく、様々な尺度があり、更に、好みは単純に1つではなく多岐にわたる好みがあると考えられる。つまり、嗜好は多次元空間上に広がりを持って分布した点の集合として定義する。当然、ユーザの嗜好にあった推薦すべきものは、多次元空間上に分布した嗜好を表す複数の点のまわりに雲のようにまとわりついていると期待される。この多次元の嗜好空間をそのままユーザに提示するのは困難である。そこで、次元削減を行い二次元のマップに射影し、それをユーザに提示する。マップ上で履歴ページの近くにある程、ユーザの嗜好に合い推薦度合いも強い。履歴ページから離れるほど強くは推薦できないが、ある履歴ページから少し離れていても別の履歴ページから近い場合もあるかもしれず、このような複雑な関係にあるページもマップで表現することでユーザに提示することが可能となる。

従来の推薦システムは、ユーザにこれはどうですか？と次々と一方的に推薦を行う。これに対してユーザはそれを選択するかどうか二者択一の判断を迫られる。そういう意味で、従来の推薦システムはテレフォンショッピング型と言える。一方、本研究はユーザの履歴から嗜好を抽出し、そこからの関係度合いを含めて推薦し、ユーザは多くの推薦されたもののなかから考えながら選択するという意味で、ウィンドウショッピング型の推薦システムと言える。推薦システムとしての効率は悪いかもしれないが、ロングテールの効果が期待できる。更に、ユーザに多くの選択肢を与えることで、ユーザが選択を楽しめるシステムであるといえる。

2.2. SOM(自己組織化マップ)

SOM(Self-organizing maps:自己組織化マップ)はT. コホネンによって提案された階層型ニューラルネットワークの一種で教師なし競学習モデルである。[3]

SOM は音声認識や画像解析といった様々な分野で

用いられており、入力されたデータの特徴に類似性があれば、マップ上の近傍領域に配置するように可視化する。本研究では T. コホネンの研究グループによって開発された SOM 用パッケージ SOM_PAK を使用した。

Web ページに対して SOM を適用した研究として波多野らの研究[4]などがある。一般に SOM は分類に使われる。波多野らも Web ページの分類に SOM を使用している。しかし、本研究では分類できるかどうかは目的でなく、多次元の特徴量を二次元に次元を減らすためだけに利用する。また、SOM を適用する Web ページも履歴ページと推薦候補ページである点でも異なる。更に、本研究ではユーザの閲覧履歴を使うだけで、ユーザに対して対話的な操作を含めて特別な操作を要求しないシステムであることも特徴である。

3. 視覚化された Web ページ推薦システム

3.1. 提案モデル

提案するシステムは閲覧履歴解析、推薦 Web ページ取得、SOM 処理、推薦 Web ページ表示の4つの処理から構成されている。

図1に処理の流れと、処理間で受け渡されるデータを示す。

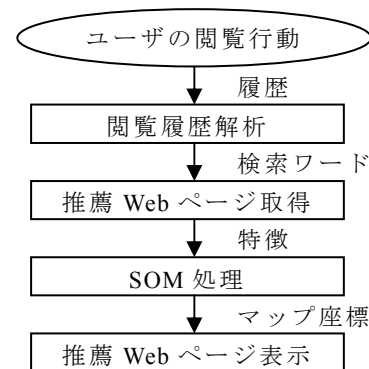


図1 Web ページ推薦の視覚化システムの概要
処理の流れと処理間で受け渡されるデータ

実験を行うためにプロトタイプシステムを構築した。次節以降で各処理部の詳細について述べる。

3.2. 閲覧履歴解析

閲覧履歴解析ではユーザが保持している履歴 URL が指す各 Web ページから検索ワードを取得する。検索ワードとは各 Web ページに出現するワードの出現回数が上位 n 個のものを指す。

図2に閲覧履歴から検索ワードを取得するまでの処理の流れを示す。

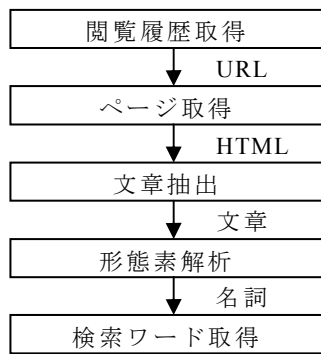


図2 閲覧履歴解析の概要

3.2.1. 文章抽出

HTML テキストから本文のみを抽出し，形態素解析に適用する文章を生成する．HTML テキストには<HTML>や<BODY>といったタグが含まれているためそれらを除く．文章内に含まれるスペース等も除去する．また履歴 Web ページの内容は SOM 用入力データ作成時にも必要となるため，タイトル，URL，ページ文書をデータとして保存しておく．

3.2.2. 形態素解析及び検索ワード取得

文章抽出によって得られたテキストに対して形態素解析を行う．形態素解析は自然言語で構成された文章を，言葉として意味を持つ最小単位である形態素に分割し，その品詞を判別する作業である．形態素解析にはオープンソースの形態素解析エンジンである MeCab を使用した．本システムでは検索ワードの品詞として名詞を用いた．しかし，名詞には接尾や非自立，また数といった直接的に Web ページの特徴となりえない語句も含まれている．そこで，「一般」及び「固有名詞」に属する語のみを対象とする．そしてそれらの文章内の出現回数をカウントした．その出現回数の上位 n 個のワードを検索ワードとして設定した．図3は文章から検索ワードを取得するまでの流れである．

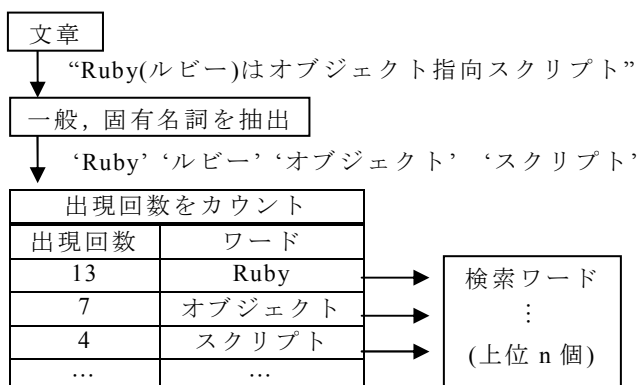


図3 文章から検索ワードを取得するまでの流れ

3.3. 推薦 Web ページ取得

閲覧履歴解析によって得られた，各履歴 Web ページの検索ワードを基に推薦 Web ページを取得する．具体的には検索エンジンを用いて得られた HTML 文書から情報を抽出する．図4に自動検索からページ情報を保存するまでの流れを示す．

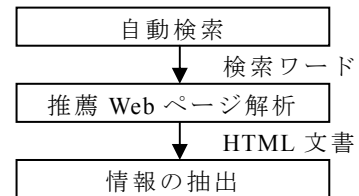


図4 推薦 Web ページ取得の概要

3.4. SOM 処理

SOM 処理ではまず履歴 Web ページ及び推薦ページから特徴ワードを抽出し，SOM 用の入力データを作成する．次に SOM の学習により履歴ページ及び推薦ページのクラスタリングを行う．その結果，推薦 Web ページ表示に使用する Web ページマップ座標を取得する．図5に入力データ作成からマップ座標取得までの流れを示す．

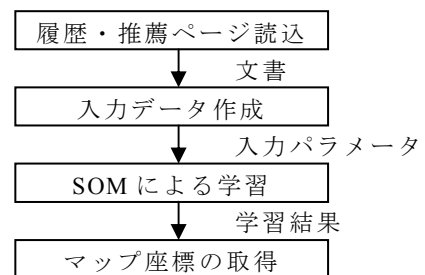


図5 SOM 処理の概要

3.4.1. 入力データの作成

SOM 用入力パラメータの詳細な作成手順について述べる．履歴ページ及び推薦ページに対して検索ワード取得処理と同様に名詞を抽出する．名詞を特徴ワードとして該当ページ文書 d における特徴ワード t の出現回数 $tf(d,t)$ を計算し，正規化する．

特徴ワードは該当するページの特徴を表しているが，その種類や個数はページ毎に異なるため，他のページに全く含まれない独立したワードを SOM に適用すると類似度計算が困難になる．そこでページ間に関連性を持たせるためにマッチングキーワードを設定する．マッチングキーワードの生成方法として，まず全てのページを対象に特徴ワード毎の $tf(d,t)$ 値の合計を計算する．次に合計 $tf(d,t)$ 値が高い上位 m 個の特徴ワードをマッチングキーワードとして設定する．図6にマッチングキーワード取得までの流れを示す．

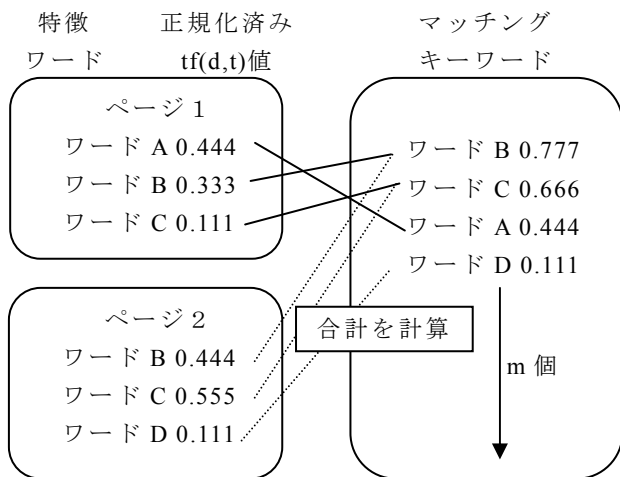


図 6 マッチングキーワードの取得方法

得られたマッチングキーワードの個数 m を SOM 入力パラメータの次元数とする．各ページが持つ特徴ワードがマッチングキーワードに含まれていれば，ページが持つ $tf(d,t)$ 値をパラメータに設定する．またマッチングキーワードを特徴ワードとしてページが保持していなければ 0 をパラメータとして設定する．図 7 に入力ベクトル設定方法を示す．

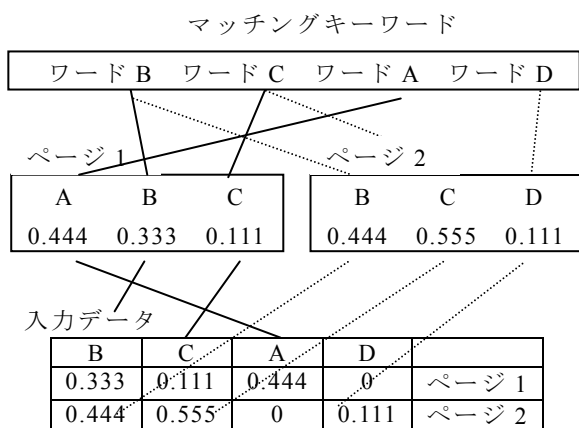


図 7 入力ベクトル設定

3.5. 推薦 Web ページマップの表示

SOM の実行によって得られた座標情報を基にマップ表示を行う．ページのタイトルをマップ上に描画した画像データ形式のマップと，実際に Web ブラウザに対応させ，履歴・推薦 Web ページへアクセス可能な推薦マップを作成した．共通する処理として SOM 処理によって得たマップ座標情報を基に対応する座標に Web ページを識別するラベルを配置する．図 8 は実際にシステムを実行して得られた推薦 Web ページマップである．マップ表示の実装は Web アプリケーション開発のフレームワークである Ruby on Rails を用いた．

マップ上に配置された Web ページタイトルにマウス

を重ねることで，対象 Web ページの概観をサムネイル画像としてポップアップ表示する．また Web ブラウザを用いることにより，専用のソフトウェアを新規にインストールする必要が無い．マップ状況に合わせてズームアップ，ズームダウンも可能である．

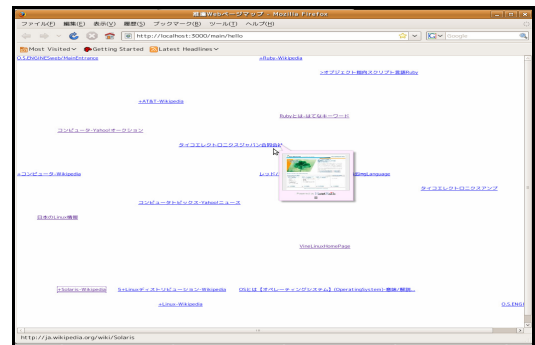


図 8 ブラウザ上のマップ表示

4. 実験

4.1. 実験設定

次節以降の実験では，サンプルとして用意した閲覧履歴（URL）を基に推薦 Web ページマップの表示を行う．マップサイズは 25×20 ，学習回数は 20000 回に設定した．適用ページ数は履歴 4 ページ，推薦ページ 12 ページである．

推薦 Web ページの自動検索には Google の検索エンジンを使用し，また履歴 Web ページと推薦 Web ページの識別手法として履歴 Web ページタイトルを「赤色」で表示し，推薦 Web ページを「青色」で表示した．

4.2. 推薦 Web ページマップの状態確認

作成したシステムを実行した結果，ユーザの興味に関連する Web ページは，興味である履歴ページに近接して配置されていることが確認できた．図 9 にシステムの実行によって得られたマップを示す．

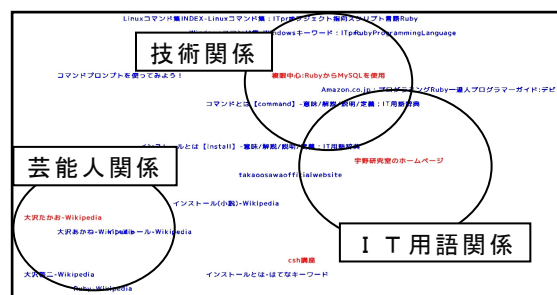


図 9 マップ上の Web ページ表示結果

このユーザは IT 用語，技術情報，芸能人といったものに興味がある．分野的に近い内容である技術関係と IT 用語はその集合も隣接しているが，大きく異なる分野である芸能人は離れて配置されている．このような関連性がマップから捉えることができるが，次に定量

的に評価を行うための実験を行った。

4.3. 距離と類似度を用いた評価実験

本実験では履歴・推薦 Web ページ間における類似度とマップ上の距離関係を見ることでマップ上の履歴・推薦 Web ページ間の関連性を評価する。

類似度は履歴ワード集合内で推薦ワード集合に含まれる単語の割合であり、履歴ページと共通したワードが多いほど、類似していると考え、値は増加する。距離は履歴から各推薦ページに対するマップ上の距離を用いる。図 10 に実験結果を示す。

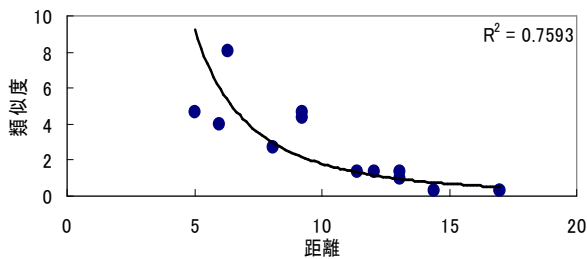


図 10 類似度と距離の関係

実験には履歴 4 ページ、推薦 12 ページを適用した。各履歴ページに対する推薦ページの類似度と距離を計算した結果、距離が 0 に近い、つまり履歴 Web ページに対して近接して配置されている推薦 Web ページは類似度が高い傾向にあった。これに対して履歴 Web ページから遠くに配置されている推薦 Web ページは類似度が低い結果になった。類似度の高い推薦 Web ページが近接して履歴 Web ページに配置されていることから、Web ページの特徴を基に的確に次元削減が行われており、履歴 Web ページ近隣には類似した Web ページが配置されていると捉えることができる。

例外として類似度は高い数値であるが、履歴からの距離が遠い推薦候補が確認された。図 11 に該当データを示す。

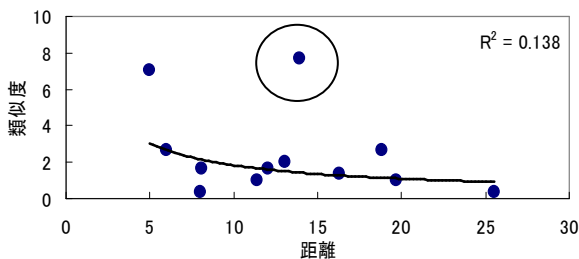


図 11 類似度と距離の関係における例外

これは SOM の適用時に履歴と推薦を区別しないこと、更に興味となる履歴ページが複数あることが原因である。ある履歴と類似度が高い推薦候補であっても、更に類似度が高い他ページが存在すれば、他の履歴ペ

ージの影響をより強く受けるため、上記のような例外データも出現する。

5. おわりに

本研究では SOM を用いることで 2 次元平面に配置された履歴ページと推薦ページの関連性を明示化することを目的とした。履歴ページと関連を持つ推薦ページをユーザに提示するシステムを実装し、評価を行った。結果、履歴 Web ページに近接する Web ページを推薦対象として扱うことで、2 次元的な推薦情報の可視化が履歴 Web ページと推薦 Web ページ間における関連性の明示化に繋がることを示した。

また、マップ上における興味名の表示や、テキスト以外のパラメータの追加、そして適用履歴の選択等を行うことにより、よりユーザの興味にあった Web ページの提供が可能であると考えられる。

参 考 文 献

- [1] 土方嘉徳”嗜好抽出と情報推薦技術(嗜好抽出・情報推薦の基礎理論”，情報処理 48(9)，957-965, 20070915
- [2] 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"
IPSJ SIG Notes 2007(65),pp 115-120.
- [3] T. コホネン著,「自己組織化マップ」, 徳高平蔵, 大藪又茂 藤村喜久朗 堀尾恵一 大北正昭 監修, シュプリンガー・フェアラーク東京株式会社,2005.
- [4] 波多野賢治, 佐野綾一, 段一為, 田中克己, "自己組織化マップと検索エンジンを用いた Web 文書の分類ビュー機構", 情報処理学会論文誌 Vol.40 No.SIG3 pp47-59, 1999.