

独立成分分析を用いた Web 閲覧履歴の解析と Web ページ推薦への応用

鶴原 翔夢[†] 高須賀清隆^{††} 丸山 一貴^{†††} 寺田 実^{††††}

[†] 電気通信大学電気通信学研究科情報通信工学専攻 〒182-8585 東京都調布市調布ヶ丘 1-5-1

^{††} 電気通信大学情報システム学研究科情報システム基盤学専攻 〒182-8585 東京都調布市調布ヶ丘 1-5-1

^{†††} 東京大学情報基盤センター 〒113-8658 東京都文京区弥生 2-11-16

^{††††} 電気通信大学電気通信学部情報通信工学科 〒182-8585 東京都調布市調布ヶ丘 1-5-1

E-mail: [†], ^{††}, ^{††††} {tom,takasuka,terada}@ice.uec.ac.jp, ^{†††}kazutaka@acm.org

あらまし 独立成分分析は、ブライント信号源分離を実現するための信号処理として発展してきた手法であり、近年では文書のベクトル空間モデルからトピックとよばれる特徴軸を見つけ出す手法としても有効であることが示されている。本研究ではそのモデルを Web 閲覧履歴解析に適用することで、閲覧履歴から閲覧行動のコンセプトと呼べるような特徴軸を検出することが可能であると考え、実際の閲覧履歴データを使って特徴軸を求めた。応用として協調フィルタリングを用いた Web ページ推薦システムに組み込み、これを用いて Web ページ推薦の評価を行ったのでその結果を報告する。

キーワード 独立成分分析, 閲覧履歴, Web ページ推薦, 協調フィルタリング

Analisis of web browsing history using Independent Component Analysis and application for web recommendation

Tom TSURUHARA[†], Kiyotaka TAKASUKA^{††}, Kazutaka MARUYAMA^{†††}, and Minoru TERADA^{††††}

[†] Graduate School of Information and Communication Engineering, The University of Electro-Communications Chofugaoka 1-5-1, Chofu, Tokyo, 182-8585 Japan

^{††} Graduate School of Information System Fundamentals, The University of Electro-Communications Chofugaoka 1-5-1, Chofu, Tokyo, 182-8585 Japan

^{†††} Information Technology Center, The University of Tokyo Yayoi 2-11-16, Bunkyo, Tokyo, 113-8685 Japan

^{††††} Department of Information and Communication Engineering, The University of Electro-Communications Chofugaoka 1-5-1, Chofu, Tokyo, 182-8585 Japan

E-mail: [†], ^{††}, ^{††††} {tom,takasuka,terada}@ice.uec.ac.jp, ^{†††}kazutaka@acm.org

Abstract In recent years, independent component analysis (ICA) is used for not only blind source separation in signal processing, but also for document topic detection in text mining. In this paper, we propose to apply ICA to the analysis of web browsing histories in web page recommendation. ICA enables us to extract the features from web browsing histories. We call the features *web browsing concepts*. We extracted the web browsing concepts from the web browsing histories of 1400+ users and integrated the similarities of the users based on the concepts into our web page recommendation system. The user evaluations are also included.

Key words Independent Component Analysis, Web Browsing History, Web Page Recommendation, Collaborative Filtering

1. はじめに

Web 上の情報量が膨大になり情報爆発と呼ばれる現象が生じている中で、ユーザにとって有用な Web ページを推薦する

システムが注目されている。このようなシステムを実現する一手法として、ユーザの閲覧履歴を用い、協調フィルタリングで推薦を生成する手法が考案されている。ユーザの閲覧履歴を用いるため特定サーバ内のページに限定されたシステムが多かつ

たが、我々はブラウザの拡張機能などを用いて閲覧履歴を収集することで、対象を Web 全体としたシステムを提案している [1] [2] [3] [4] .

協調フィルタリングでは複数ユーザ間の類似度を計算し、類似度の高いユーザの好むページを推薦する。あるページの閲覧回数をそのページへの評価値と見なし、ユーザプロフィールとして用いる場合、単純なコサイン尺度などを用いて類似度を定義すると、同一の Web ページを見ていないユーザ同士の類似度は 0 となってしまう。Web 全体を対象とする場合には、異なるユーザが同一のページを見ている可能性は低く、このような類似度では嗜好の似たユーザが見つけれないことが多い。これは Web ページの全体量に対して、あるユーザが閲覧するページ数が非常に少ないことに起因するデータのスパース性から生じる問題である。

スパース性にかかる問題に対しては、文書分類・検索の分野で様々な手法が考案されている。クラスタリングを用いる手法、潜在意味解析 (Latent Semantic Analysis, LSA) などがよく用いられているが、近年では独立成分分析 (Independent Component Analysis, ICA) を用いたトピック検出を利用する手法も提案されている。どの手法も、データから特徴軸を抽出し、それらの軸でデータを再構成することでスパース性を解消しているという点で共通している。

本研究では文書分類で用いられるこれらの手法を閲覧履歴に適用することを提案する。特に ICA に関して、閲覧履歴データに適用する場合の ICA モデルの解釈について述べる。さらにこれを使って求められた特徴空間上での類似度を推薦システムに応用する。推薦システムは [1] で提案しているシステムを基にしている。ユーザプロファイリングには Web ページの閲覧回数を用い、協調フィルタリングを利用したシステムである。1434 名分の実データを用いた実験を通してこれらの手法の評価を行う。

本稿の構成は以下の通りである。まず 2 章において関連研究について述べた後、3 章で ICA のモデルとそれを閲覧履歴データに適用する場合の解釈について述べる。4 章では具体的な推薦アルゴリズムを示し、5 章で実データを用いた評価実験の結果を示す。最後にまとめと今後の課題について述べる。

2. 関連研究

文書のベクトル空間モデル上で ICA を用いた研究としては、Isbell らによる [8] や Kolenda らによる [7] がある。従来用いられていた LSA によって求めた特徴軸よりも ICA で求めた特徴軸の方が、文書分類に有効な場合があることを示している。濱本らは [10] で、トピック検出に対する性能に関して、クラスタリングによる手法、LSA、ICA による手法を比較している。さまざまなトピックを含んでいるような文書において、ICA が有効であることを示している。本研究では Web 閲覧履歴データをベクトル空間モデルと同様にベクトルで表現し、これに ICA を適用することで、閲覧履歴の特徴軸を求める。さらにこの特徴軸を使ってユーザ間の類似度を測り、Web ページ推薦へ応用する。

	USER1	USER2	USERn
URL1	h_{11}	h_{12}	h_{1n}
URL2	h_{21}	h_{22}	
.....	
URLm	h_{m1}			h_{mn}

図 1 履歴行列 (h_{ij} はユーザ j の URL i に対する重みである)

特定の Web サイトに限定しない Web 推薦システムとしては、Zhu らの "Web ICLite" [6] や、丹羽らのソーシャルブックマークを用いたシステム [5] がある。本研究の推薦システムのベースとしている [1] のシステムでは、これらの研究と比較して以下のような特徴を備えている。

- ユーザはブラウザの拡張機能をインストールするだけで、システムを使用できる。
- コンテンツの中身を使用せず URL のみを用いて推薦を生成するため、動画や Flash などリッチコンテンツを推薦することが可能である。

3. 独立成分分析の閲覧履歴への適用

3.1 ICA モデル

独立成分分析 (ICA) は信号処理の分野でブライント信号源分離という問題を解くために開発された統計手法である。

複数の信号源が存在しているところに、複数のマイクを設置して音声を観測する状況を考える。簡単のため、信号源の数とマイクの数と同じであるとし、この数を k と置く。原信号を $S = (s_1(t), \dots, s_k(t))^T$ とする。 T は転置を表す。 $s_i(t)$ はある時刻 t における信号の振幅値である。観測される信号 $X = (x_1(t), \dots, x_k(t))^T$ が原信号の線形和であると考えれば $X = AS$ という関係が成り立つ。ここで $A = (a_1, \dots, a_k)$ は混合行列と呼ばれる。ある時刻における信号の振幅値を確率変数の実現値と考えると、 X, S は確率ベクトルであることができる。ICA では S の各要素の値が統計的に独立に決まるという仮定から、 X だけを元にして A と S を推定する。ICA を解くためのアルゴリズムとしては Hyvärinen らによる fastICA [11] がよく用いられる。本研究でもこの手法を使用する。

3.2 閲覧履歴に適用する際の解釈

閲覧履歴に ICA モデルを適用するために、各ユーザの履歴をベクトルで表現する。対象とするデータ集合のうちのユニークな URL の数を m とすると、ベクトルは m 次元となる。各要素の値はあるユーザの各 URL に対する、評価値となる。この値は単純に閲覧回数としても良いが、極端に閲覧回数が多い場合の影響を抑える目的で、ここでは $h_{ij} = \log(1 + f_{ij})$ とする。ここで f_{ij} はユーザ j が URL i を閲覧した回数である。このベクトルを以降では履歴ベクトルと呼ぶ。 n 人分のユーザの閲覧ベクトルを並べると図 1 のような行列 H が構成できる。この行列を履歴行列と呼ぶ。

H を ICA モデル式の X と対応づけ、 $H = AS$ というモデル式を考える。ユーザのインデックスが音声モデルにおける時刻に対応し、各 URL がマイクの役割を担う。あるユーザの閲覧履歴は様々な興味が反映されているはずである。ICA モデルでは A の各列ベクトルが、この興味を表現したベクトルであり、これを S の各要素で重み付け重ね合わせた結果、履歴ベクトルが構成されていると見る。このとき、各興味への重みである S の各要素が独立に決まっていると仮定することになる。

A の各列は、データ空間上の特徴軸であると見ることもでき、ベクトル空間モデルに適用する場合には、トピックと呼ばれているものに相当する。上では直感的に理解しやすい興味という語を用いたが、この特徴軸は必ずしも我々が興味という語に対して抱くイメージと合致するとは限らないと考えられる。ICA モデルは、これらの特徴軸に対する重みである S の各要素が統計的に独立に決まるということを仮定するのみである。しかし、この特徴軸はなんらかの典型的な閲覧パターンを表現しているものと考え、以降ではこれを閲覧コンセプトベクトルと呼ぶ。

4. 推薦アルゴリズム

本研究の Web 推薦システムの具体的な推薦生成手順の概略は以下のようにになっている。

(ステップ 1) ブラウザの拡張機能を用いてユーザから閲覧履歴データを収集する。

(ステップ 2) 集まったデータから履歴行列 H を構成しこれを解析して特徴軸を求める。さらに、求めた特徴軸上に各ユーザの履歴ベクトルを射影し、この上で類似度を計算する。

(ステップ 3) 求めた類似度を使って、推薦対象となる URL すべてに対してページスコアを計算する。

(ステップ 4) ページスコアの高いものを推薦ページとする。

これらの手順のうち、特徴軸を求める方法、ページスコアの計算について詳しく説明する。

4.1 特徴軸の抽出

特徴軸の抽出に ICA を用いる。ICA で求められる閲覧コンセプトベクトルがそのまま特徴軸となる。単純な ICA のモデルでは、音源数とマイク数が等しいことを仮定している。今、音源数は特徴軸の数、マイク数はユニーク URL の数に対応しているため、一般には後者の方が大きな値をとる。そのためまずはデータの次元圧縮を行う。これには H の特異値分解を用いる。

$$H = UDV^T \quad (1)$$

ここで、 U, V は直交行列、 D は特異値を対角要素に持つ対角行列である。これらのうち大きな特異値に対応するベクトル k 本だけを取り出したものをそれぞれ U_k, D_k, V_k とすると、 H を次元圧縮した X は

$$X = U_k^T H = D_k V_k^T \quad (2)$$

となる。これを ICA の入力として

$$X = AS \quad (3)$$

となる A, S を求める。 $U_k A$ が特徴軸、 S がその軸上への各履歴ベクトルの写像である。 S の各列は特徴空間上でのユーザプロフィールとなるので、このコサイン尺度をもって各ユーザ間の類似度とする。 S の各列ベクトルを正規化して S' とすれば、全ユーザ間の類似度を並べた類似度行列 R が次の用に計算できる。

$$R = S'^T S' \quad (4)$$

R は対称行列で、要素 r_{ij} は i 番目のユーザと j 番目のユーザの類似度となる。

4.2 ページスコアの計算

求めた類似度を用いて、推薦対象となる URL のそれぞれに対してスコアを計算する。先ほどの類似度行列と同じように全ユーザに対する各ページのスコアを並べた、ページスコア行列 P を考え以下のように計算する。要素 p_{ij} は i 番目の URL の j 番目のユーザに対するスコアである。

$$P = C\hat{H}R \quad (5)$$

$\hat{H} = U_k D_k V_k^T$ であり、 C は多くのユーザが閲覧している有名サイトのスコアを抑えるための係数を対角要素に持つ対角行列である^(注1)。 C の対角要素は $c_{ii} = f(n_i)$ とする。 n_i は i 番目の URL を閲覧した人数である。 f は閲覧人数に対する関数で、ここでは次のような関数とした

$$f(x) = \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (6)$$

パラメータの値はそれぞれ $\mu = 2.0, \sigma^2 = 50.0$ とした。合計で 2 人が閲覧した URL のスコアを最も高く評価し、15 人が閲覧した URL にはその 0.1 倍程度の重みを付けることになる。ある一人だけしか閲覧していないページは、特殊性が高すぎると考えて 2 人が閲覧した URL のスコアが最高となるようにした。

5. 評価実験

本来は、ユーザがブラウザの拡張機能を使って送信した閲覧履歴データを用いて実験を行う必要があるが、特異値分解や ICA を行うためには 1000 単位のユーザデータを使うことが望ましく、現状ではそれほどの被験者を集めることができていない。そこで、電気通信大学の対外接続部に設置されたスニッファの出力データを用いて、閲覧履歴データを代替することとし、これを使って評価実験を行った。

実験に用いたデータは 2007 年 12 月 18 日のデータ一日分である。ブラウザの拡張機能を用いて収集した閲覧履歴には、ユーザが明示的にクリックしたページの URL のみが現れる。しかしスニッファの出力データには、ページに付随してリクエストされる画像やスクリプトなどが含まれているため、これらをできる限り除去する必要がある。クローラなど人手によらないリクエストも含まれていることも考慮する必要がある。以降

(注1): 多くの人が利用するサイトは、ユーザも既知である可能性が高く、このようなサイトが推薦として出現することは好ましくないと考えている。

表 1 閲覧コンセプトの例 (左列の数値はベクトルの成分の値である．これは URL の閲覧コンセプトへの影響力を表す．)

0.0103675	http://www.yahoo.co.jp/
0.00671671	http://fxfeeds.mozilla-japan.org/rss20.xml
0.00440347	http://ja.fxfeeds.mozilla.com/ja/firefox/headlines.xml
0.00403104	http://jp.msn.com/
0.00379516	http://pa.yahoo.co.jp/bc?
0.00375099	http://search.yahoo.co.jp/chfs
0.0032352	http://www3.asahi.com/rss/index.rdf
0.101427	http://headlines.yahoo.co.jp/cm/comment_articlejs.php
0.0904269	http://headlines.yahoo.co.jp/lib/ctlFs_1.0.php?
0.0484616	http://headlines.yahoo.co.jp/hl?c=ent&t=1
0.0401901	http://headlines.yahoo.co.jp/hl?c=soci&t=1
0.0392855	http://www.yahoo.co.jp/
0.0392533	http://headlines.yahoo.co.jp/hl
0.0346179	http://headlines.yahoo.co.jp/hl?c=soci&t=1&p=1
0.0248492	http://ja.fxfeeds.mozilla.com/ja/firefox/headlines.xml
0.0248035	http://news.google.co.jp/nwshp?tab=wn
0.0179172	http://news.google.co.jp/nwshp?tab=wn&ned=jp&topic=w
0.014643	http://meta.wikimedia.org/w/index.php?title=Special:NoticeLoader&action=raw
0.011917	http://news.google.co.jp/nwshp?tab=wn&ned=jp&topic=t
0.0116047	http://news.google.co.jp/nwshp?tab=wn&ned=jp&topic=p
0.0105498	http://www.google.co.jp/webhp?hl=ja

ではこのようなリクエストの URL を不要 URL と言う．今回は以下の手順で不要 URL の除去を試みた．

(1) 使用しているユーザエージェントの種類の総数が 50 種類を越えるユーザは NAT を使用しているローカルネットワークの内側からのアクセスと考え除去．

(2) 広く普及しているブラウザとは異なるユーザエージェントによるリクエストを除去．

(3) 連続するリクエストシーケンスの中で，前のアクセスから 1 秒以下の間隔でリクエストされたものを除去．

(4) リクエストされたファイル名の拡張子が gif, png, jpeg, jpg, js, css, ico であるものを除去．

(5) 人の明示的なクリックによってアクセスすることがないと考えられるサイトや，個人情報を含むと考えられるサイトのリストを著者の主観で作成し，このリストに含まれるサイトへのアクセスを除去．

これらの処理を行った結果，ユニークユーザ数 1434 人，ユニーク URL 数 89776 件の閲覧履歴データを得た．

5.1 閲覧コンセプトの抽出

上記のデータを用いて実際に閲覧コンセプトベクトルを求めた．ICA の前処理として，特異値分解を用いて閲覧履歴データを 100 次元まで圧縮してある．そのため閲覧コンセプトに相当するベクトルは計 100 本得られる．そのうちの一例を表 1 に示す．表に示したのは，閲覧コンセプトベクトルのうち，値の大きな次元に対応する URL である．

表 1 上段を見るとポータルサイトやニュースサイトの RSS フィードなどが現れている．これらの中には一般的なブラウザが立ち上がったときに自動で読み込まれる URL が多く含まれ

ている．表 1 中段ではニュース記事の URL が多く現れている．また表 1 下段ではニュース記事でもサイトが異なる URL 群が現れている．

表 1 上段の中にあるの”http://pa.yahoo.co.jp/bc?”などは 1 ドットの画像に対する URL であり，前述した不要 URL に含まれるはずのものである．このように本来のブラウザの拡張機能を用いて収集したデータには現れないはずのデータも影響を及ぼしていることが分かる．

その他の閲覧コンセプトを見ても，ブラウザ起動時に読み込まれる URL など多くの人が閲覧する URL が強く現れている例が多く見られた．実データにより求めた閲覧コンセプトは，我々が普段，“興味”と言って連想する，“スポーツ”や“プログラミング”といった分かりやすい形の興味わかるようなベクトルにはなっていないと言える．しかし，“ニュースを見る”，“動画サイトを見る”，といった典型的な閲覧行動を表現したようなベクトルになっているような例は複数見られた．

5.2 ソーシャルブックマークを用いた評価

ICA で求めた特徴軸を使ったユーザ間の類似度がどのような性質を有しているのかは直感的には明らかではない．

ところで，ユーザの興味をユーザが直接表現しているものとして，ソーシャルブックマークにおけるタグがある．タグとは，あるブックマークに対してユーザが，そのページの性質や分類を表すキーワードを関連づけたものである．あるユーザのすべてのブックマークに対するタグを集めると，これはそのユーザの興味を反映したキーワードの集合になっているはずである．

あるユーザが付けたタグの集合は，閲覧履歴と同じようにベクトルで表現できる．ユニークなタグの総数が l 個であるとす

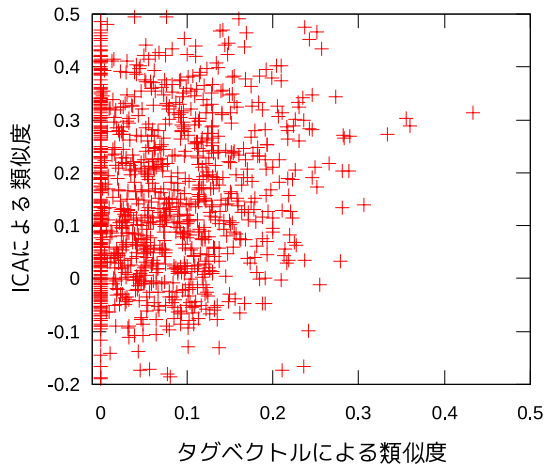


図2 タグベクトルによる類似度とICAによる類似度の比較

れば、このベクトルは l 次元のベクトルとなる。ベクトルの各次元はタグひとつに対応し、その値はユーザのそのタグに対する評価値となる。これをタグベクトルと呼ぶことにする。タグはユーザの主観によって付けられたものであるから、タグベクトル同士のコサイン尺度は、ユーザ間の類似度として有用であると考えられる。

ICAによる類似度とタグベクトルをつかった類似度の相関を調べる目的で以下のことを行った。

(1) はてなブックマーク [9] から 1024 名分のブックマークデータ及び各々のブックマークに関連づけられたタグを取得。

(2) ユーザごとにタグベクトルを構成し、全ユーザ間の類似度を計算。

(3) ユーザのブックマークデータをそのユーザの閲覧履歴と見なし、ICA を用いて類似度を計算。

ユーザの閲覧履歴を使った類似度が計算できることが望ましいが、はてなブックマークを利用しているユーザの閲覧履歴を知ることが不可能なため、ブックマークを閲覧履歴と置き換えることとした。

図2に示したのは、タグベクトルによる類似度とICAによる類似度の相関関係を散布図で表現したものである。無作為に選んだある一人のユーザに対して、全ユーザとの類似度を2種類の手法でもとめ、タグベクトルによる類似度の値を横軸に、ICAによる類似度を縦軸に取り、プロットしたものである。二つの類似度の間に相関があれば散布図は直線上に分布するはずである。

図を見るとタグベクトルによる類似度が低いユーザに対して、ICAによる類似度が高くなっている場合が多く、二つの類似度の間に相関関係は認められない。

タグは、必ずしもそのページの内容を表しているとは限らず、ユーザの感想や顔文字などが含まれていることもある。またタグの表記ゆれの問題もあるため、ユーザの趣味が似ていても、必ずしもタグベクトル同士の類似度が高くなるとは限らない。今回はこのことを考慮しなかったが、ICAによる類似度はタグによる類似度とは異なっていると言える。

5.3 ユーザによる推薦システムの主観評価

学生8名を被験者として、実際に各人の閲覧履歴から推薦を生成し、評価を行ってもらった。評価は、

- 自分の普段の興味に沿う内容であるか
- 面白かったか

という項目に、それぞれ4段階の数値で答えてもらった。比較のため3種類の手法で推薦を10件ずつ生成し、これをランダムな順序に並び替えて被験者に提示した。各手法は前節で述べたアルゴリズムを基本としているが、類似度計算の方法などを変更している。具体的には以下のようになっている。

手法1(ICA) 前節で述べたアルゴリズムによる手法。

手法2(LSA) $\hat{H} = U_k D_k V_k^T$ の各列をノルムが1となるように正規化し、 \hat{H}' とする。これを用いて類似度行列を置き換え、 $R = \hat{H}'^T \hat{H}'$ としたものの。

手法3(COS) H の各成分 $h_{ij} = f_{ij}$ とし、さらに各列を正規化したものを H' とする。これを用いて類似度行列を $R = H'^T H'$ で置き換え、さらにページスコア行列 $P = CHR$ としたものの。

一つ目の方法は、これまで見てきたICAによる手法である。二つ目は類似度の計算をLSAで求めた特徴軸上で行う手法、三つ目の手法では、ページスコアの計算に \hat{H} を用いることができないため、 H でこれを代用する。

本実験で用いているデータには不要URLが含まれており、推薦結果としてこれらのURLが出ることもある。実運用する際にはこのような推薦が出る事はおこりえないため不適な推薦であると考え、ユーザに提示する前に除去した。

ユーザ評価の結果を図3, 4に示す。各手法で生成した10件の推薦に対する評価値の平均値を被験者ごとにプロットしたものである。図3は、興味に沿う内容であったか、図4は面白かったかという問いに対する評価値で、数値が大きいほど良い評価である。横軸の数値は被験者を識別するための番号である。全被験者の評価値の平均は表2に示す。

手法ごとの評価値の差について wilcoxon の符号付順位検定を行ったが、有意水準5%でも各手法間に有意な差は認められなかった。

このことからICAによる特徴軸を使った類似度によるWebページ推薦は、推薦システム全体として有効に機能しているとは言えない。しかしユーザによってはICAによる推薦が最も良いと評価している場合もあることから、条件によっては有効な推薦を生成することがあると言える。

ICAではユーザの閲覧履歴が複数の閲覧コンセプトの重ね合わせであると考えられる。今回の実験に用いたデータは1日分の履歴データであったため、閲覧履歴に複数のパターンが現れていることが少なく、これが原因でICAが有効に働かなかったのではないかと考えられる。

コサイン尺度による類似度の問題点は、類似度0となるユーザが多く、類似ユーザが発見できないことがあるということであったが、今回の被験者はすべてコサイン尺度で類似ユーザを発見することができていた。コサイン尺度と他の手法との差があまり見られなかったのは、このことにも起因していると思わ

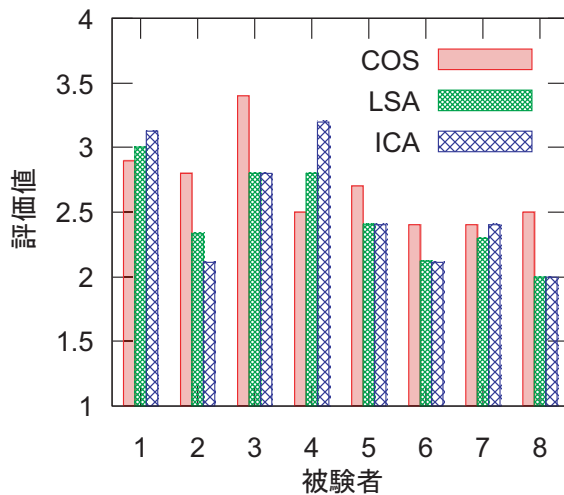


図3 ユーザ評価結果 (興味に沿う内容であったか)

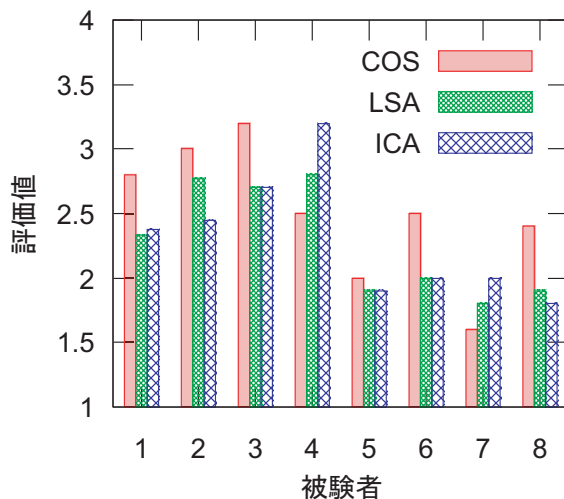


図4 ユーザ評価結果 (面白かったか)

表2 ユーザ評価結果の平均値

	COS	LSA	ICA
興味に沿う内容であったか	2.70	2.47	2.51
面白かったか	2.50	2.28	2.30

れる。

6. まとめと今後の課題

Web 閲覧履歴に独立成分分析を適用することで、ユーザの閲覧履歴の閲覧コンセプトと呼べるような特徴軸を抽出することを提案し、実データからこの特徴軸の例を示した。またこの特徴軸を利用してユーザ間の類似度を測る協調フィルタリングをベースにした推薦システムを提案した。ユーザの主観評価実験の結果、すべてのユーザに対しての有効性を示すことはできなかったが、ユーザによっては良い推薦を生成することを示した。

今後の課題としては以下のようなものが挙げられる。

(実データの質の向上) ユーザ実験に用いるデータは、ブラウザの拡張機能を使って収集することが望ましいが、多くのユーザデータを得る事は難しいため、スニッファの出力のような

データを用いる必要がある。しかし、これには前述の不要 URL の問題などがあり、統計処理にも影響を及ぼしていると考えられる。スニッファの出力データから不要 URL をより精度よく除去する手法の開発が必要である。

(被験者を増やしての評価実験) 今回の実験の被験者は学生 8 名であったが、結果からは ICA を用いた手法と、比較に用いた手法の間に有意な差は見られなかった。より多くの被験者を対象にした評価実験を行う必要がある。

(長い期間の履歴データを用いた実験) 今回の実験では 1 日分の履歴データを使用したため、閲覧履歴に複数興味が見れにくかったと考えられる。より長い期間のデータを用いた実験が必要である。

現在は閲覧コンセプトと考えられる特徴軸を求めたあとで、すべての特徴軸を利用して類似度を計算している。ICA によって求められた特徴軸はベクトル空間モデルにおいてはトピックと呼ばれる。閲覧履歴に適用した場合にも、ユーザの閲覧行動から、なんらかの典型的な閲覧パターンを抽出したものであると考えられる。そのためこの閲覧コンセプトをユーザに複数提示し、ユーザに選択させることで、そのパターンに応じた推薦が可能になると考えられる。閲覧コンセプトは各次元が URL に対応するベクトルとして得られるため、これをそのままユーザに提示してもユーザは選択の基準を得にくい。これをわかりやすく提示する手法が必要である。

謝 辞

有用なユーザデータを提供して頂いた電気通信大学情報基盤センターの関係者の方々に感謝致します。

文 献

- [1] 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,” DBWS2007, vol.107, No.131, pp.115-120, 2007.
- [2] K.Maruyama, K.Takasuka, Y.Yagihara, S.Machida, Y.Shiirai, M.Terada, “Real-time Discovery of Currently and Heavily Viewed Web Pages,” proc. of WEBIST 2006, pp.352-259, 2006.
- [3] K.Takasuka, M.Terada, K.Maruyama, “Web Page Recommendation by URL-based Collaborative Filtering,” proc. of WEBIST 2007, pp.447-450, 2007.
- [4] 高須賀清隆, 白井雄一郎, 丸山一貴, 寺田実, “閲覧履歴を共有するウェブブラウザ,” FIT2005, pp.355-356, 2005.
- [5] 丹羽智史, 土肥拓生, 本位田真一, “Folksonomy マイニングに基づく Web ページ推薦システム,” 情処学論, vol.47, No.5, pp.1382-1392, 2006.
- [6] R.Greiner, T.Zhu, G.Haubl, K.Jewell, “A Trustable Recommender System for Web Content,” Beyond Personalization 2005, pp.83-88, 2005.
- [7] T.Kolenda, L.L.Hansen, “Independent Components in Text,” Advances in Neural Information Processing Systems, vol. 13, pp.235-256, 2000.
- [8] C.L.Isbell, P.Viola, “Restructuring Sparse High Dimensional Data for Effective Retrieval,” Advances in Neural Information Processing Systems, vol.11, pp.480-486, 1998.
- [9] はてなブックマーク. <http://b.hatena.ne.jp/>.
- [10] 濱本雅史, 北川博之, Jia-Yu Pan, and Christos Faloutsos, “独立成分分析を用いたテキストデータからのトピック検出,” DEWS2004, 3-B-04, 2004.
- [11] A.Hyvärinen, J.Karhunen, E.Oja, 詳解独立成分分析, 電機大学出版局, 東京, 2005.