携帯向けパーソナルオンラインニュース配信のための視聴/非視聴履歴に基づく嗜好クラスタ管理手法

大槻 一博† 服部 元‡ 星野 春男† 松本 一則‡ 菅谷 史昭‡

†NHK 放送技術研究所 〒157-8510 東京都世田谷区砧 1-10-11
‡KDDI 研究所 〒356-8502 埼玉県ふじみ野市大原 2-1-15

E-mail: †{otsuki.k-ek, hoshino.h-ii}@nhk.or.jp, ‡{gen, matsu, fsugaya}@kddilabs.jp

あらまし 視聴履歴に基づきユーザの嗜好に適応したニュース記事の推薦を行う，携帯向けオンラインニュース配信サービスについて検討している．視聴履歴には様々な嗜好情報が混在しているものと考え，これまでに，視聴した記事の履歴から関連のある内容を持つ記事群である嗜好クラスタを生成し，嗜好クラスタ毎の重み付けを行うことで複数の嗜好情報を適応的に管理する手法を提案している．この手法においては，視聴履歴のみを使用して嗜好クラスタを生成していたが，はっきりとしたユーザの嗜好が視聴履歴に明確に現れない場合には推薦精度が低下する課題があった．そこで本稿では，推薦精度を向上するため，ユーザが視聴しなかった記事の履歴である非視聴履歴も利用した 2 通りの嗜好クラスタ管理手法を提案した．推薦精度の評価実験の結果，提案手法の平均精度がこれまでの手法の平均精度よりも向上していることを確認した．また，代表的なクラスタリング手法であるウォード法とクラスタリングの精度に関する評価実験を行い，ウォード法よりも望ましいクラスタリング手法であることを示した．さらに，ユーザの全履歴中の非視聴履歴の割合に応じた適応的な嗜好クラスタ管理手法の切り替えが，推薦精度向上に有効であることを示した．

キーワード オンラインニュース配信，非視聴履歴，嗜好クラスタ管理手法

A Preference Cluster Management Method based on User Access/Non-viewed Logs for Personal Online News Delivery toward Portable Terminals

Kazuhiro OTSUKI† Gen HATTORI‡ Haruo HOSHINO†

Kazunori MATSUMOTO‡ and Fumiaki SUGAYA‡

†NHK Science & Technical Research Laboratories 1-10-11 Kinuta, Setagaya-ku, Tokyo, 157-8510 Japan
‡KDDI R&D Laboratories Inc. 2-1-15 Ohara, Fujimino-shi, Saitama, 356-8502 Japan

E-mail: †{otsuki.k-ek, hoshino.h-ii}@nhk.or.jp, ‡{gen, matsu, fsugaya}@kddilabs.jp

Abstract We examine online news delivery toward portable terminals that recommends the news article that adapts to a user preference based on user access logs. Based on the hypothesis that multiple user interests can be obtained from user access logs, we proposed the new method of the preference management using the following technique; extraction of multiple user preferences by clustering articles in user access logs, and application of a weight to each cluster. In the conventional method, there was a problem to which the accuracy of the recommendation decreased when the user preference did not clearly appear to the user access logs. To improve the accuracy of recommendation, this paper describes two kinds of preference cluster management techniques that used non-viewed logs in addition to user access logs that are the conventional method only uses. Our experiments concerning accuracy of recommendation resulted that use of non-viewed logs gives higher averaged accuracy than that in case of use of user access logs only, and that the value of the information entropy of the proposed method is smaller than that of Ward's method. These results indicate the cluster generation by the proposed method is preferable clustering to the Ward's method. We also confirmed that adaptive switch of method for cluster generation as a function of the ratio of non-viewed logs among all logs of the user is effective to the improvement of accuracy of recommendation.

Keyword Online news delivery, Non-viewed logs, Preference cluster management method

1. はじめに

昨今，通信インフラの急速な普及により，Web 上での情報発信は当たり前のように利用されるようになり，オンラインニュース配信は新聞各社,ポータルサイト，放送事業者など様々な事業者が情報サービスを提供するようになった．さらにそれらのサービスにおいて，全てのユーザに同じ情報を提示するだけでなく，個人がカスタマイズ可能なオンラインニュース配信サービスも開始されている．このことから，ユーザの興味や関心などの嗜好情報に応じた情報が選択されて提示されることが望まれているといえる．例えば，Google ニュース[1]においては，表示するページをカスタマイズ可能であり，ログインすることで検索履歴に基づいたお勧めニュースを表示することが可能である．また My Yahoo![2]においては，関心のあるニュースだけをキーワードやカテゴリでカスタマイズして表示でき，freshEYE NewsWatch [3]においては,トピックを指定することでカスタマイズが可能である．このように，各ユーザが所望するニュースを簡単，確実に視聴したいという要望が存在する．

我々は，このようなオンラインニュース配信においても特に，携帯端末向けのサービスに着目する．ここで, 携帯端末の限られた処理能力, 画面表示や UI(User Interface)では情報の一覧性に乏しく，また嗜好情報をユーザに入力させることは困難である.そのため,ユーザ所望の情報が自動で簡単に表示されることが望ましい．また，例えば野球と経済の記事を好んで視聴するなど，ユーザの嗜好は 1 つの分野ではなく複数分野の嗜好が混在するため,その入力は煩雑となってしまう．そこで我々は，ユーザには極力手間を掛けさせない方法として，ユーザの視聴行動の表れである視聴した記事の履歴（以降，視聴履歴と呼ぶ）に基づく方法について検討している．視聴履歴に基づくユーザ嗜好の抽出に関して，嗜好情報を１つの特徴ベクトルで表現しているものは，ユーザの複数分野の嗜好の平均となるため特徴が消えてしまう課題がある．また複数の嗜好として扱っているものであっても，それらを対等に扱った場合は，配信される記事数が少ないマイナー分野の嗜好情報は記事数が多いメジャー分野の嗜好情報に埋もれてしまう課題がある．また，あらかじめ決められたジャンルやキーワードを用いて複数の嗜好を扱う手法は，新たなジャンルやキーワードに対応する嗜好を扱うことが困難である課題がある．

以上のことを鑑み，我々はこれまでに，視聴履歴に基づきユーザの複数の嗜好を抽出して嗜好クラスタを生成し，嗜好クラスタの重みを個別に管理する嗜好クラスタ管理手法を提案している[4]．ここで，嗜好クラスタとは，ユーザが関心を持った記事群である視聴履歴から，関連のある内容を持つ記事群同士をまとめたものである．ここでは，ユーザの興味の判断対象になる記事のタイトルを利用して嗜好クラスタを生成している．これにより，新たなキーワードの発見が可能となり，あらかじめ決められたジャンルやキーワードにとどまらない複数嗜好を抽出できることを示し，これまで抽出できなかったマイナー分野の嗜好に対して有効に働く可能性を示した.しかしながら,ユーザがはっきりした嗜好を持っているにも拘らず，視聴履歴からだけではその嗜好が明確に読み取れない場合には，推薦精度が下がってしまうという課題があった．

そこで本稿では，ユーザの嗜好を正確に反映した嗜好クラスタ生成を実現して推薦精度を向上するため，視聴履歴だけではなく,視聴しなかった記事の履歴(以降，非視聴履歴と呼ぶ）を利用する手法を 2 手法提案する．これにより，ユーザの複数嗜好に高精度に適応するニュース記事推薦サービスの実現を目指す．

本稿の構成を述べる．まず，2.では目標とするサービスの概要を述べる．3.ではこれまでに提案した嗜好情報の抽出・管理手法について述べる．4.では視聴/非視聴履歴を利用した嗜好クラスタ管理手法について述べる．5.では実装したシステムで行った提案手法の推薦精度に関する評価実験の結果および考察を述べる．6.で関連研究について述べ，7.でまとめる．

2. 目標とするサービスの概要

本研究の目標は，ユーザの嗜好に応じた携帯端末向けの情報提供サービスの実現である．本稿では具体例として，携帯向けのオンラインニュース配信サービスを検討する．目標とするサービスの全体概要を図 1に示す．

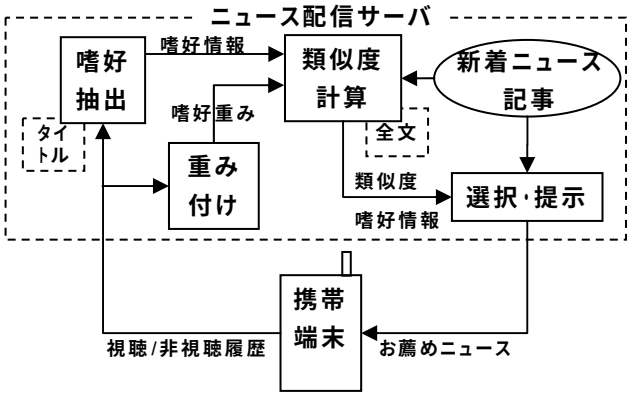

図 1 サービスの全体概要

携帯端末向けのサービスであるため，本サービスではユーザには嗜好情報の入力や登録などの煩雑な操作を強要しない．ここでは，携帯端末の操作に基づく視聴履歴を自動的に取得することとする．ニュース配信

サーバは，各ユーザの視聴履歴中の各記事のタイトルからキーワードを抽出し，嗜好情報を生成する．同時に嗜好情報への重み付けを行う．次に，この嗜好情報と新着ニュース記事の全文を用いて類似度を計算し，類似度と嗜好情報に基づきユーザにお薦めのニュースを選択して携帯端末に提示する．これらの手順を繰り返すことにより嗜好情報が蓄積されるため，推薦精度が向上する．

オンラインニュース配信のソースとしては，NHKオンライン(http://www.nhk.or.jp/)におけるニュースを対象とする．このソースにおける１件の記事は，タイトル，ヘッドライン，詳細記事からなる．図 2にその一例を示す.各タイトルは最大 13 文字から構成されており，スペースを用いることで助詞を省くなど，記事の内容をコンパクトに表現されている．従って，ユーザはこのタイトルを見て，内容を自身の知識で類推し，その記事を読むかどうかの最初の判断をすることができる．すなわち，タイトルを選択しヘッドラインを読んだ記事がユーザの興味のある記事と判断することができる．よって，視聴履歴としては，ユーザが記事タイトルを選択してヘッドラインまで視聴したときに初めて視聴した履歴を記録するものとする．

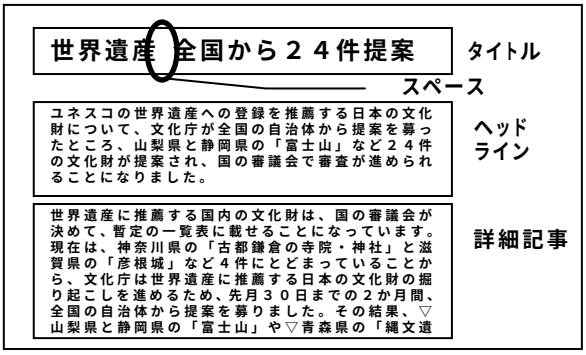

図 2 ニュース記事の一例

3. 視聴履歴を利用した嗜好クラスタ管理手法

3.1. 嗜好クラスタ管理手法

2.で述べたサービスを実現するため，これまで，視聴履歴に基づきユーザの複数の嗜好を抽出して嗜好クラスタを生成し，嗜好クラスタの重みを個別に管理する嗜好クラスタ管理手法を提案している[4]．ここでは，記事のタイトルから茶筅[5]を用いた形態素解析で抽出したキーワードを利用して嗜好クラスタを生成している．嗜好クラスタの特徴量として，各嗜好クラスタに属するニュース記事から TF-IDF[6]に基づく嗜好クラスタベクトルを算出する．算出方法の詳細は付録 1 に示す．また，嗜好クラスタに属する各ニュース記事が記録された時刻から現在時刻までの時間差によって定まる嗜好クラスタの重みを定義し，嗜好クラスタベクトルと合わせて用いることで嗜好情報を管理する．

ユーザがニュース配信サービスにアクセスすると，嗜好クラスタ管理手法から得られた嗜好情報に基づきニュース記事の推薦を行う．まず，新着記事iの特徴ベクトルD_iと，嗜好クラスタCの嗜好クラスタベクトルQ_Cとのコサイン距離を計算する．次に，嗜好クラスタCが持つ重みW_Cを乗じてその記事と嗜好クラスタとの類似度$sim(Q_C,D_i)$を算出する．式 1 に示すように，$sim(Q_C,D_i)$の最大値を記事ベクトルD_iとユーザとの類似度$S(D_i)$と定義し，$S(D_i)$の大きい順に配信ニュース記事を推薦する．

$$S(D_i)=\max_C\left[\, sim(Q_C,D_i)=W_C\frac{Q_C\cdot D_i}{|Q_C|*|D_i|}\right] \quad (1)$$

3.2. 課題

例えば,「イチロー選手」の記事だけには興味があるユーザが「イチロー選手」の記事の視聴履歴を残している状態を仮定する．この場合，3.1.の手法では，他の類似記事に対する情報がないことから,「イチロー選手」以外の記事には興味が無いという嗜好を得ることができない．そのため,「イチロー選手」の記事と同じジャンルに含まれる大リーグの「松井選手」の記事が新着記事にあった場合には，その記事を推薦してしまう可能性があり，さらにそれを修正することができないため何度も提示してしまう課題がある．また，例えば台風の季節のニュースには,「台風」の記事が多数配信される可能性がある．その場合，一度「台風」の記事を視聴してしまうと,「台風」の記事ばかりが推薦されてしまい，他のニュースが殆ど表示されない状態になってしまう課題がある．

以上の検討結果より，以下の 2 つの課題が挙げられる．

(課題 1) ユーザの嗜好がはっきりしているにも拘らず，視聴した履歴に記録されたニュース記事からだけではその嗜好が明確に読み取れない場合には，推薦精度が下がる可能性がある．

(課題 2) $S(D_i)$の大きい順にニュース記事を提示するだけでは，新たに配信すべき記事の内容によっては，記事同士の内容が類似したものばかりが提示される可能性がある．

4. 視聴/非視聴履歴を利用した嗜好クラスタ管理手法の提案

3.2.で示した課題 1 を解決するため，視聴履歴のみでなく非視聴履歴も利用した嗜好クラスタ管理手法を 2 手法提案する．なお，課題 2 の解決については，嗜好クラスタ管理手法に基づく階層型の提示をすることで満足できるため，5.3.の実装例で具体的に述べる．

4.1. 方式 1：嗜好/非嗜好クラスタの独立管理手法

視聴履歴を用いて嗜好クラスタを生成する手法と同様の手法で，非視聴履歴を用いて非嗜好クラスタを生成し，それぞれの嗜好クラスタを独立に管理する．それぞれの嗜好クラスタに対し，式 2 に示す忘却関数により求めた値に基づく重み付けを行う．この忘却関数は,古い履歴の影響を抑えるような重み付けができる．

現在時刻から記事iが記録された時刻までの時間差によって定まる嗜好クラスタCの重みW'_Cは, 現在時刻τ，記事iが記録された時刻T_i，減衰期間T，嗜好クラスタに含まれる記事数mを用いて，

$$W'_C = \sum_{i=1}^{m} W_0 \exp\left(-\lambda\left(\tau - T_i \diagup T\right)\right) \quad (2)$$

と定義する．なお，$\lambda(0<\lambda<1)$は減衰期間 T の増加に対する減衰の度合いを表す忘却定数である．ニュース記事の推薦は次の手順で行う．

(1) 新着記事に対し，全ての嗜好クラスタおよび全ての非嗜好クラスタとの類似度を算出する．ここでは，式 1 のW_C の代わりに，式 2 で定義したW'_C を用いる．

(2) (1)で算出した類似度の最大値を新着記事とユーザとの類似度$S(D_i)$とする．類似度が最大値となるクラスタが非嗜好クラスタの場合には，類似度$S(D_i)$をマイナスの値とする．

(3) 全ての新着記事でユーザの嗜好クラスタとの類似度を求め，類似度$S(D_i)$が大きなニュース記事から順に推薦する．

上記の手法により，ユーザの嗜好を反映していない非嗜好クラスタに近いニュース記事ほど下位の候補となるニュース記事の優先順位が得られる．すなわち非視聴履歴の嗜好情報に強く影響を受ける手法であるといえる．

4.2. 方式 2：嗜好/非嗜好クラスタの統合管理手法

視聴履歴並びに非視聴履歴を統合し，ひとつの履歴として嗜好クラスタを生成し，嗜好情報を管理する．それぞれのクラスタに対し，式 3 に示す閲覧率に基づく重み付けを行う．

クラスタ中に含まれる記事の閲覧率と時間順に基づいた嗜好クラスタCの重みW''_Cは，嗜好クラスタ中の視聴した記事数n，嗜好クラスタに含まれる記事数m，視聴した記事iの嗜好クラスタ中の最新からの時間順位P_iを用いて，

$$W''_C = \frac{n}{m}\sum_{i=1}^{n}\frac{1}{p_i} \quad (3)$$

と定義する．ニュース記事の推薦は 3.と同様の方法で行う．ただし，式 1 のW_C の代わりに，式 3 で定義したW''_C を用いる．

この手法により，含まれる記事数の少ない嗜好クラスタでも閲覧率の高い嗜好クラスタに類似している記事であれば上位の候補とすることができる．

4.3. 非視聴履歴の記録

提案する手法は，ユーザが視聴しなかった記事が，ユーザが興味を持たなかった記事であるという前提に基づいている．非視聴履歴を残すニュース記事の対象を新着ニュース記事全てとすると，本当に興味が無かったのか，単に他の理由で視聴しなかったのかが判別できず，結果として推薦精度が下がってしまうことになる．よって，非視聴履歴としては，サービスにアクセスして最初に提示された記事項目であるにも関わらず，サービス終了時までに視聴されなかった記事を，視聴しなかった履歴として記録するものとする．こうすることで，システムが推薦したにも関わらず，ユーザが視聴しなかった記事を記録でき，効率良く嗜好情報を生成することが可能となる．

5. 評価実験

提案手法の有効性を確認するため，提案手法で推薦したニュース記事が，ユーザの嗜好に合致するかどうかの推薦精度の評価を行う．ここでは 3 つの評価実験を行った．

5.1. 実験手順

それぞれの実験手順を以下に述べる．

（実験 1)「タイトルからのキーワード抽出の妥当性」

提案手法は, NHK オンラインのニュース記事のタイトルが，記事内容の重要な部分を要約して作成されていることを前提とし，ニュース記事のタイトルのみで嗜好クラスタを生成している．そのため，この記事タイトルからのキーワード抽出手法が妥当なものか検証する必要がある.そこで,タイトルに含まれているキーワードにおける，記事作成者により独自にアレンジされた単語の有無を確認する．ここでは，2005 年 4 月から蓄積している約 10000 件のニュース記事から無作為に抽出した 200 件の記事に対し，タイトルから抽出したキーワードがその記事のヘッドラインに含まれているか否かを調べる．

（実験 2)「非視聴履歴利用の有効性評価」

視聴/非視聴履歴の両方を利用した嗜好クラスタ管理手法の有効性を評価する．ここでは視聴履歴のみを利用した場合を比較対象とし，提案手法のうち，方式 1 と評価した．比較評価の指標として，システムが提示したニュース記事の平均精度[7]を用いる．平均精度の詳細を付録 2 に示す．平均精度の値が大きいほど，ユーザの嗜好に合う正解記事が上位の候補となってい

ると判断できるため，この値がユーザ嗜好の適応度合いを示しているといえる．本実験では，上位 10 位内に推薦した記事に対する平均精度を用いる．また，忘却定数λ=0.9，W_0=1 とし，実験の手順を以下に示す．

(1) 4 人のユーザ A,B,C,D に対し，1 週間に 1 度，100 件の記事を視聴させる．各ユーザは，興味のある記事および興味のない記事にそれぞれ○印と×印を付ける．これを 16 週繰り返す．

(2) 最新の 100 件に含まれる○印の記事を正解記事集合とする．それ以前の 15 週分の記事を集計し，視聴履歴および非視聴履歴を作成する．

(3) 作成した視聴履歴から嗜好クラスタを生成し，また非視聴履歴から非嗜好クラスタを生成する．

(4) 嗜好クラスタのみを利用した方法と提案手法とでニュース記事の推薦を行う．

(5) (4)で推薦したニュース記事の平均精度を計算する．

(6) 減衰期間 T を変化させて各手法の平均精度を比較する．

（実験 3)「提案手法の嗜好クラスタ生成精度の評価」

本稿で提案している 2 つの嗜好クラスタ管理手法の嗜好クラスタ生成精度について評価する．ここでは，代表的なクラスタリング手法であるウォード法[9]と比較した．嗜好クラスタ生成の評価指標として，クラスタ毎の情報エントロピー[8]を用いる．情報エントロピーを算出することで，各クラスタ内の情報集約度を測定することができる．情報エントロピーの算出方法を付録 3 に示した．本実験では，ユーザが記事に興味がある/ないの 2 択であるので,式 A･4 の正解集合の個数 H は 2 とする．対象とするニュース記事は 100 件を 1 セットとして 3 セット用意した．実験 2 において差が顕著に現れたユーザＢとＣについて，各手法で生成した嗜好クラスタの情報エントロピーを比較する．

5.2. 実験結果と考察

それぞれの実験における結果と考察を以下に述べる．

（実験結果 1）

表 1に結果を示す．200 件の記事のタイトルから得られたキーワードのうち，84%はその記事のヘッドライン中にキーワードがそのまま同一表現で含まれていた.また,同一表現が含まれていない場合については，省略表現や別表現など，意味は同じである語彙は必ず含まれていた．なお，タイトルに用いられるキーワードは，基本的に文字数の少ない方が使用されている．また，これらの表現の違いは，タイトルのみでクラスタ生成を行っている提案方式には影響しない．以上より，提案方式のタイトルからのキーワード抽出は，妥当であると考えられる．

表 1 タイトルから得られたキーワードの詳細

パターン	件数	例
同一表現が含まれる	168 件 (84%)	--
省略表現である	7件 (3.5%)	デジカメ⇔デジタルカメラ、東証⇔東京証券取引所
別表現である	13 件 (6.5%)	首相⇔総理大臣、準々決勝⇔ベスト 8、米⇔アメリカ
名詞と動詞の表現違い	12 件 (6%)	容疑⇔疑い、削減⇔減らす
対応表現が含まれない	0 件 (0%)	--
合計	200 件 (100%)	--

（実験結果 2）

図 3に結果を示す．4 人のユーザ全てにおいて，非嗜好クラスタ利用の場合に，嗜好クラスタのみの場合と比較して，同じか高い平均精度を示している．非視聴履歴の利用によって，平均精度の向上に寄与しているといえる．

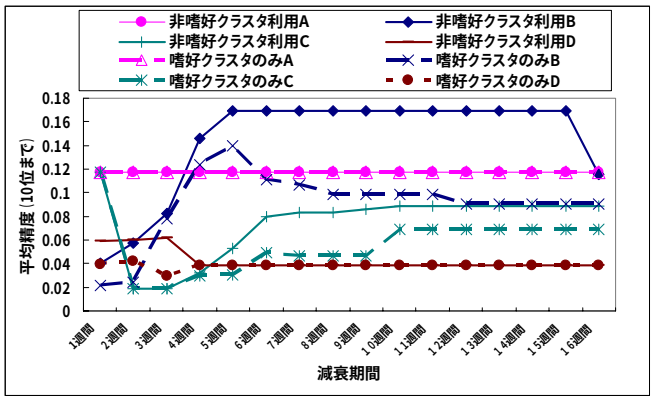

図 3 非嗜好クラスタ利用有無の比較

また，6 週目以降の平均精度はほとんど一定で高い値をほぼ維持している．これは減衰期間 T が長ければ長いほど良いことを示しており，最終的には実装するシステムおいてニュース配信サーバが処理可能な値を，減衰期間として指定すれば良いといえる．なお，ユーザＡの全期間に渡って差が無い原因は，最新の 100 件中の正解記事数が 85 件と多く,嗜好が広範囲に渡っているためである．また，ユーザＢの 16 週目において非嗜好クラスタ利用の場合に精度が落ちている原因は，16 週目の重みが上がったことによる変化が，嗜好クラスタにはあったが，非嗜好クラスタには無く，その影響により，それまでは非嗜好クラスタに類似しているという理由で順位を下げていたユーザの嗜好には合わない記事が，嗜好クラスタに類似するようになり上位の候補のまま残ってしまったためである.このように，現在の嗜好が，広範囲に渡っている場合には精度の差が出ず，過去からの嗜好とは一致しない嗜好に変化し

ている場合には,稀に精度が低くなる可能性もあるが,嗜好クラスタのみの場合と比べて，同じかそれより高い精度を保っており，問題は無いと考える．

（実験結果 3）

図 4に結果を示す．各提案手法とウォード法について，情報エントロピーを求めてグラフにした．また，各ユーザの視聴記事数に対する非視聴履歴の割合を表 2に示す．

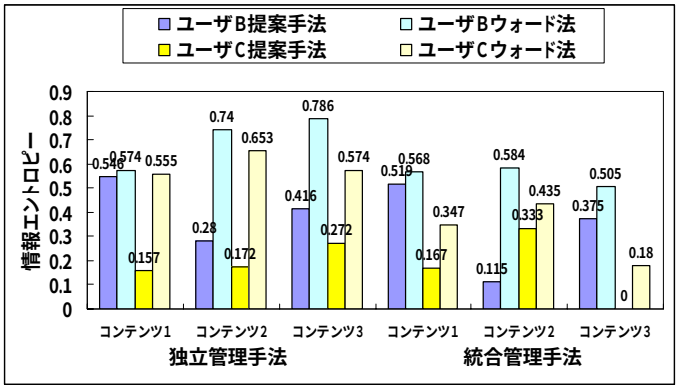

図 4 各手法による情報エントロピーの比較

表 2 視聴記事数に対する非視聴履歴の割合

ユーザ	A	B	C	D
割合	56%	53%	78%	71%

どちらの提案手法も，比較手法であるウォード法より情報エントロピーの値が小さくなっており，最大で 0.48 の差が見られた．これにより，提案方式の方がウォード法よりも高精度なクラスタ生成が実現できていると言える．

また実験は，表 2に示すように，非視聴履歴の割合が両極端にあるユーザ B と C に対して行ったが，ユーザ B においては，統合管理による手法の方が独立管理による手法に比べて，全てのコンテンツにおいて情報エントロピーの値が低いことから，統合管理手法の方が高精度なクラスタリングであるという結果となった．これは，ユーザ B の非視聴履歴の割合が 53%であり，他のユーザと比較して履歴中の視聴履歴の割合が高いため，閲覧率を利用することのクラスタリングへの効果が大きいことを表していると考えられる．よって，履歴中の視聴履歴の割合が高いユーザに対しては統合管理による手法が有効といえる．一方のユーザ C においては，コンテンツによって多少のばらつきはあるものの，独立管理による手法の方が統合管理による手法よりも望ましいクラスタリングであるという結果となった．これは，ユーザ C の非視聴履歴の割合が 78%と，履歴中の非視聴履歴の割合が高いことで，非嗜好クラスタによるクラスタリング効果が大きいことを表していると考えられる．よって，履歴中の非視聴履歴の割合が高いユーザには独立管理による手法が有効であるといえる．以上より，ユーザの全履歴中の非視聴履歴の割合に応じて統合管理による手法と独立管理による手法を適応的に切り替えることで,嗜好/非嗜好クラスタ生成の精度を向上させる可能性を示した．

5.3. 提案手法の実装例

3.2.の課題 2 に対する解決方法として，嗜好クラスタを応用した階層型の提示方法による実装例を示す．

想定するサービスとして，ユーザの嗜好に合った内容であれば，同一内容のニュースであっても多数提示するのではなく，ユーザの嗜好に合う記事においても幅広い内容の記事を閲覧することを望ましいと考えられる．携帯端末の限られた画面内に多種の記事が提示されることを可能とする提示方法として，本稿の提案手法を利用できる．すなわち，提案手法により生成した嗜好クラスタに基づき類似記事群を生成し，2 段階で提示する階層型提示が可能となる．階層型提示の画面例を図 5に示す．

最初の提示では，嗜好クラスタのキーワードとそれぞれのクラスタのトップ記事のタイトルを提示する．次に，嗜好クラスタのキーワードを選択すれば，その嗜好クラスタに類似している記事群のタイトルを提示し，トップ記事のタイトルを選択すればその記事のヘッドラインと詳細記事を提示する．ここでトップ記事とは，ある嗜好クラスタに類似する記事中で類似度の最も高い記事とする．

図 5 階層型提示の画面例

このような各嗜好クラスタによる階層型の提示方法により，同一内容の記事が連続して提示されることを避けながら，加えて多種の記事を提示できるようになり，ユーザが嗜好に合う記事を閲覧する機会を増やすことが可能となる．

6. 関連研究

ユーザの嗜好情報に適応させるサービスを行うためには，ユーザの嗜好を取得・収集・抽出すること，すなわち，ユーザプロファイリング技術と，得られた嗜好に基づいてユーザに適応した情報を選び出すこと，すなわち，情報フィルタリング技術が必要である．情報フィルタリング技術としては，コンテンツに基づくフィルタリングと協調フィルタリングの 2 種類がある．前者は，推薦する情報の内容に基づき情報の選択を行い，後者は，同じ嗜好を持ったユーザ群を発見し，そのユーザ群が好む情報を選択する[10]．本研究では前者の手法を用いる．いずれの手法においても，ユーザの嗜好を把握するためのユーザプロファイリング技術は必要となり，ユーザの嗜好を得ようという目的の研究は多数存在する．

嗜好情報を得る間接的な手法として，閲覧履歴や操作履歴といったユーザの視聴行動から抽出する方法がある．閲覧履歴利用による Web ページ提示システムとして，利用者の閲覧履歴に基づき興味の分類体系を動的に構築し，利用者の興味に基づき自律的に情報を分類し融合して提示するという特徴を持つ Web の情報融合システムについて提案している研究[11]がある．ユーザ嗜好を複数の嗜好として扱ってはいるが，すべての嗜好を並列に扱っており，複数の嗜好を重みを持たせて扱っている我々とは異なる．

文章や情報の重要度に忘却の概念を取り入れている研究としては，オンライン環境では，ユーザは一般に新規性の高い文章に対して興味を有することを考慮し，忘却の概念を導入した文書クラスタリング手法を提案しているもの[12]や，トピック分析のために文章の影響力の時間による逓減モデルを用いている研究[13]がある．より新しい文章ほどユーザが興味を持つという点や，より自然な減衰関数として指数関数を用いている点などは我々と同一の考え方であるが，あくまでも文章のクラスタリングを考えており，ユーザ嗜好の抽出を行っていない．また，ユーザ嗜好の抽出を行い，興味の減衰項として指数関数を用いている研究[14][15]があるが，ユーザ嗜好を単一に扱っており，複数嗜好からどのように文章を推薦するかの言及がない点が我々と異なる．

Web のデータマイニングや閲覧履歴の視覚化などにおいて，Web アクセス履歴の解析がいくつか行われている．各ユーザが閲覧した類似 Web サイトをクラスタリングする際に，ウィンドウタイトルから単語を抽出して用いることで，Web アクセスログデータからユーザの興味の遷移パターンを抽出するデータマイニング手法を提案している研究[16]や，リンク選択時の履歴としてタイトルから得られるキーワード群を用いることで，Web ブラウジング履歴から，キーワードと Web ページを表すアイコンを 2 次元空間上に配置して時系列に提示する手法を提案している研究[17]がある．そのページの内容が代表されているものとして，タイトルからキーワードを抽出する考え方は我々と同一であるが，Web のデータマイニング手法であり，その興味データからどのようにして Web を推薦するかの言及がない点が我々と異なる．

クラスタリングすることにより嗜好を得る手法として，自己組織化マップを用いる研究がある．出力セルのベクトル成分の中で最も値の大きな単語をそのセルのキーワードとし，それにより分類マップの領域形成をしている研究[18]や，多くの履歴が分類されたクラスタにユーザの嗜好が集中していると考える研究[19]があるが,明示的な適合フィードバックを必要とせず,マイナー分野の嗜好情報をも得ようとする我々とは観点が異なる．

7. おわりに

携帯向けオンラインニュース配信を対象としてユーザに適応した情報サービスについて検討した．これまでの手法では，はっきりとしたユーザの嗜好が視聴履歴に明確に現れない場合には推薦の精度が低下してしまう課題があった．その解決手法として，視聴した履歴に加えて，視聴しなかった履歴も利用する嗜好クラスタ管理手法を提案した．

推薦の精度に関する評価実験の結果，クラスタ生成時のタイトルからのキーワード抽出は妥当であること，提案手法の平均精度がこれまでの手法の平均精度よりも向上していることを確認した．また，代表的なクラスタ生成手法であるウォード法とクラスタ生成の精度に関する評価実験を行い，提案手法の嗜好クラスタ内の情報エントロピーがウォード法の場合よりも低い値が得られ，ウォード法よりも望ましいクラスタリング手法であることを示した．さらに，ユーザの全履歴中の非視聴履歴の割合に応じた適応的な嗜好クラスタ管理手法の切り替えがユーザ適応サービスに有効となる可能性を示した．最後に提案手法を利用した実装例を示し，多種の記事を効率的に提示できる階層型提示が可能になることを示した．

文　献

[1] Googleニュース http://news.google.co.jp

[2] My Yahoo! http://my.yahoo.co.jp/

[3] freshEYE NewsWatch http://news.fresheye.com/top/

[4] 大槻一博,服部元,帆足啓一郎,星野春男,菅谷史昭,“携帯向けオンラインニュース配信のための視聴履歴に基づく嗜好クラスタ管理手法の検討,”電子情報通信学会ヒューマンコミュニケーショングループ WI2 研究会資料, pp.113-118, 2006.7

[5] 松本裕治,北内啓,山下達雄,平野善隆,松田寛,高岡一馬,浅原正幸,“日本語形態素解析システム『茶筌』version 2.3.3 使用説明書,”奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座, 2003

[6] G.Salton and M.J.McGill, “Introduction to Modern Information Retrieval”, McGraw-Hill, 1983

[7] 酒井哲也,“よりよい検索システム実現のために,”情報処理学会誌, Vol.47, No.2, pp.147-158, 2006.2

[8] Michael Steinbach, George Karypis, Vipin Kumar, “A Comparison of Document Clustering Techniques,” KDD Workshop on Text Mining, 2000

[9] 神嶌敏弘,“データマイニング分野のクラスタリング手法(1)－クラスタリングを使ってみよう！－,”人工知能学会誌, Vol.18, No.1, pp.59-65, 2003

[10] 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,”人工知能学会誌, Vol.19, No.3, pp.365-372, 2004

[11] 河合由起子,官上大輔,田中克己,“興味と好みに基づく複数 Web ページの情報融合・提示システムの検討,”電子情報通信学会データ工学研究専門委員会第 15 回データ工学ワークショップ(DEWS2004) 2-C-01

[12] 石川佳治,北川博之,“忘却の概念に基づくクラスタリング手法の改良方式,”日本データベース学会 Letters, Vol.2, No.3, pp.53-56, 2003

[13] 崔春花,北川博之,“到着頻度と関連性を考慮した文書ストリームのトピック分析,”電子情報通信学会データ工学研究専門委員会　第 15 回データ工学ワークショップ（DEWS2004）3-C-01, 2004

[14] 佐竹聡,川島英之,今井倫太,“ニュースコンテンツ提示ロボットにおけるユーザ興味を考慮したコンテンツ選択手法,”電子情報通信学会データ工学研究会 技術研究報告, Vol.105, No.171, pp.119-124, 2005.7

[15] Sugiyama,K.,Hatano,K.,and Yoshikawa,M. “Adaptive Web Search Based on User Profile Constructed without Any Effort from Users,” Proc. of WWW’02, pp.675-684, 2004

[16] 山田和明,中小路久美代,上田完次,“Web ユーザの行動履歴解析のためのデータマイニング,”電子情報通信学会ヒューマンコミュニケーショングループ WI2 研究会資料, pp.59-64, 2005.9

[17] 村上晴美,平田高志,“Web におけるリンク選択行動からユーザの時系列の興味空間を作成するシステム,”日本認知科学会テクニカルレポート, JCSS-TR-47, pp.1-12, 2003

[18] 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（TOD1), pp.47-59, 1999

[19] 芥子ら,“ディジタル情報家電のインタフェースエージェント技術の開発,”シャープ技報,第 77 号, pp.15-20, 2000

付　録

1. 嗜好クラスタベクトル

嗜好クラスタに含まれる記事から，嗜好クラスタの代表ベクトルとして，嗜好クラスタベクトルを得る．ニュースの各記事の統計量は，TF-IDF[6]によるベクトル表現を用いる．嗜好クラスタに含まれるニュースkの記事のベクトルD_kを全て総和して，それを嗜好クラスタの代表ベクトルとする．すなわち，嗜好クラスタベクトルQ_Cは，嗜好クラスタに含まれる記事数をmとして，

$$Q_C = \sum_{k=1}^{m} D_k \qquad (A\cdot 1)$$

と定義する．

2. 平均精度[7]

平均精度とは，「各正解記事が提示された項目順位での精度を全ての正解記事で平均したもの」である．すなわち，全正解数を R，システムが出力した検索結果のサイズを L，検索結果の第 r 位における文書が正解であるとき 1，正解でないとき 0 となるフラグ $I(r)$，第 r 位までに含まれる正解の個数を $count(r)$ $(\leqq r)$で表すと，第 r 位における精度 $P(r)$は，

$$P(r) = \frac{count(r)}{r} \qquad (A\cdot 2)$$

と書くことができ，このときの平均精度は，

$$\text{平均精度} = \frac{\sum_{r=1}^{L} I(r)P(r)}{R} \qquad (A\cdot 3)$$

と定義される．

3. 情報エントロピー[8]

情報エントロピーとは，「事象の不確かさの相対値」である．すなわち，正解集合K_h $(1\leqq h\leqq H)$にクラスタC_k $(1\leqq k\leqq L)$の文章が属する確率$P(K_h|C_k)$を考えると，各クラスタのエントロピーは，

$$E_k = -\sum_{h=1}^{H} P(K_h \mid C_k)\log P(K_h \mid C_k) \qquad (A\cdot 4)$$

と定義できる．クラスタC_kに含まれる文章数をn_k，K_hとC_kとに共通の文章数をn_{hk}とすると，確率$P(K_h|C_k)$は，

$$P(K_h \mid C_k) = \frac{n_{hk}}{n_k} \qquad (A\cdot 5)$$

と推定され，もしクラスタC_k中の文章が様々な正解集合に属するならば，エントロピーE_kの値は増加する．したがってE_kの値が小さいほど，望ましいクラスタリングであると解釈できる．クラスタリングの結果全体に対しては，クラスタに関する全エントロピー$E(C|K)$，文章数による重み付け平均として，

$$E(C \mid K) = \sum_{k=1}^{L} \frac{n_k}{N} E_k \qquad (A\cdot 6)$$

と定義される．