

携帯向け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

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

大槻 一博^{†1} 服部 元^{†2} 星野 春男^{†3}
松本 一則^{†4} 菅谷 史昭^{†5}

Kazuhiro OTSUKI Gen HATTORI
Haruo HOSHINO Kazunori MATSUMOTO
Fumiaki SUGAYA

視聴履歴に基づきユーザの嗜好に適應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應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ワ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、ワ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應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可能性を示した。

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.

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行われている。また、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。

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。携帯端末の限られた処理能力、画面表示やUI(User Interface)では情報の一覽性に乏しく、また嗜好情報をユーザに入力させることは困難である。そこで、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。これまでに、視聴履歴に基づきユーザの複数の嗜好と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た[1]。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。この手法では、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、ユーザがはっきりと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。また、評価実験を行い、本方式の有効性の検証を行う。

2. 目標とするサービスの概要と従来手法

2.1 サービス概要

本研究の目標は、ユーザの嗜好に応じた携帯向けニュース配信によりユーザの満足する情報提供サービスを実現することである。目標とするサービスのための提案システムの構

†1 正会員 NHK放送技術研究所 otsuki.k-ek@nhk.or.jp
†2 正会員 株式会社KDDI研究所 gen@kddilabs.jp
†3 非会員 NHK放送技術研究所 hoshino.h-ii@nhk.or.jp
†4 非会員 株式会社KDDI研究所 matsu@kddilabs.jp
†5 非会員 株式会社KDDI研究所 fsugaya@kddilabs.jp

成を図1に示す。

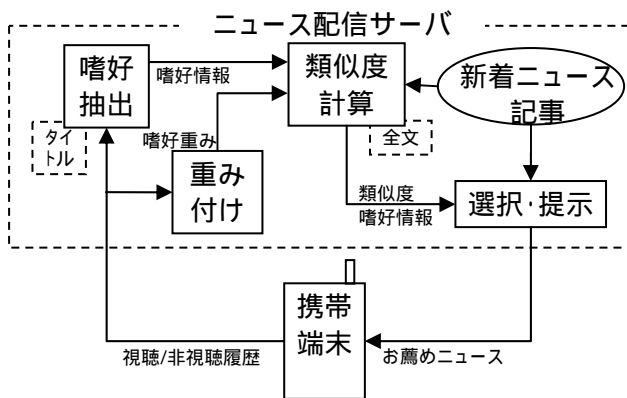


図1 提案システムの構成

Fig.1 Proposal System Configuration

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。

2.2 従来手法と課題

2.1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と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[1]。ここでは、記事のタイトルから茶筌[2]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のTF-IDF[3]による特徴ベクトルを全て総和した嗜好クラスタベクトルを算出する。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。

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

しかしながら、例えば、「イチロー選手」の記事だけで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、上記の従来手法では、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ことが考えられる。すなわち、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という課題がある。

3. 視聴/非視聴を利用した提案手法

3.1 方式1：独立管理手法

2.2で述べた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、それぞれのクラスタを独立に管理する。それぞれの嗜好/非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 、クラスタに含まれる記事数 m を用いて、

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

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/非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からの嗜好情報に強く影響を受ける手法であるといえる。

3.2 方式2：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2.2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。

3.3 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は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。

4. 評価実験

4.1 実験手順

(実験1)「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[4]を用いる。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」であり,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 $\lambda=0.9$, $W_0=1$ 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,毎週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2)「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ワード法[6]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[5]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。本実験では,ユーザが記事に興味があるかないかの2択であるので,正解集合の個数は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1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。

4.2 実験結果と考察

(実験結果1)

図2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。すなわち,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。

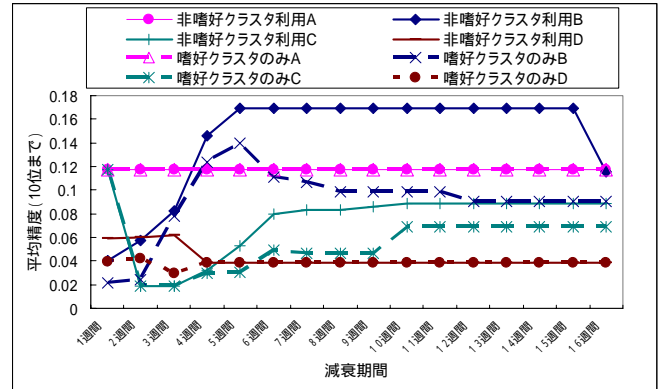


図2 非嗜好クラスタ利用有無の比較

Fig.2 Comparison of Non-preference Cluster Use Existence

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に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。

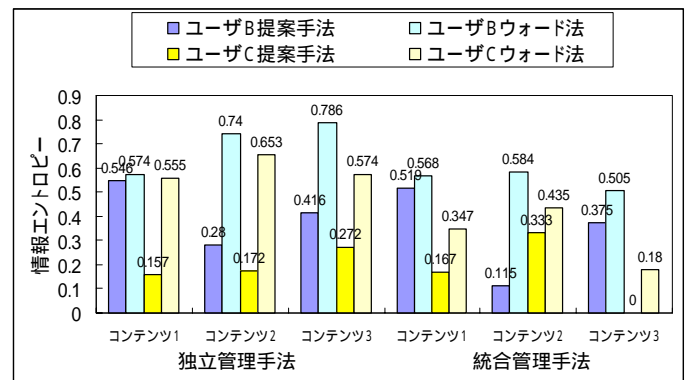


図3 各手法による情報エントロピーの比較

Fig.3 Comparison of Information Entropy by Each Method

(実験結果2)

図3に結果を示す。各提案手法とワード法について、情報エントロピーを求めてグラフにした。また、各ユーザの全履歴中の非視聴履歴の割合を表1に示す。

表1 全履歴中の非視聴履歴の割合
Table 1 Ratio of Non-viewed Logs in All Logs

ユーザ	A	B	C	D
割合	56%	53%	78%	71%

どちらの提案手法も、比較手法であるワ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、提案方式の方がワ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1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が行え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いことを表していると考えられる。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。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、よりユーザの嗜好に合った嗜好/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. まとめ

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提供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、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ワ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ワード法の場合よりも低い値が得られ、ワ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。

[文献]

[1] 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

- [2] 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
- [3] G.Salton and M.J.McGill, “Introduction to Modern Information Retrieval”, McGraw-Hill, 1983
- [4] 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
- [5] Michael Steinbach, George Karypis, Vipin Kumar, “A Comparison of Document Clustering Techniques,” KDD Workshop on Text Mining, 2000
- [6] 神鳥敏弘, “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

大槻 一博 Kazuhiro OTSUKI

平成7年東工大・工・情報工学卒業。平成9年同大大学院修士課程修了。同年NHK入社。現在、放送技術研究所システム(新サービス)研究員。この間、データ放送方式、多重化方式の研究開発に従事。平成16年度映像情報メディア学会鈴木記念奨励賞受賞。映像情報メディア学会、日本データベース学会各会員。

服部 元 Gen HATTORI

平成8年神戸大・工・電気電子工学卒業。平成10年同大大学院修士課程修了。同年国際電信電話(株)(現KDDI(株))入社。現在、(株)KDDI 研究所知能メディアグループ研究主査。この間、ネットワーク管理、ITS、ソフトウェアエージェント、Webコンテンツマイニングの研究開発に従事。平成15年電子情報通信学会学術奨励賞受賞。電子情報通信学会、情報処理学会、日本データベース学会各会員

星野 春男 Haruo HOSHINO

昭和63年早大・理工・電子通信卒業。同年NHK入社。現在、放送技術研究所システム(新サービス)主任研究員。この間、立体テレビ方式、情報家電システム、ネットワーク映像通信システムの研究に従事。電子情報通信学会、映像情報メディア学会各会員。

松本 一則 Kazunori MATSUMOTO

昭和59年京都大・工・情報工学卒業。昭和61年同大大学院修士課程修了。同年国際電信電話(株)(現KDDI(株))入社。現在、(株)KDDI 研究所知能メディアグループ主任研究員。この間、マルチメディア検索、コンテンツ配信の研究開発に従事。平成10年人工知能学会研究奨励賞、平成12年度電子情報通信学会論文賞を各受賞。電子情報通信学会、情報処理学会各会員。

菅谷 史昭 Fumiaki SUGAYA

昭和57年東北大・工・通信工学卒業。昭和59年同大大学院修士課程修了。同年国際電信電話(株)(現KDDI(株))入社。現在、(株)KDDI 研究所知能メディアグループリーダー。この間、情報検索、e-Learning、音声翻訳評価の研究開発に従事。平成3年電子情報通信学会学術奨励賞受賞。平成18年電子情報通信学会情報システムソサイエティ論文賞受賞。電子情報通信学会、日本音響学会、情報処理学会各会員。工博。