RF-006

類似ユーザ群を用いた非有益嗜好の抽出手法の提案
Extracting Useless Information from Similar-preference Users

近藤 司 ‡　伊藤 真也 ‡　原田 史子 †　島川 博光 †
Tsukasa Kondo　Masaya Ito　Fumiko Harada　Hiromitsu Shimakawa

1. はじめに

WWW 上に存在する膨大な web ページの中から，ユーザが必要な情報を適切に取捨選択するのは困難であるため，情報推薦の技術が研究されてきた．

既存の情報推薦手法の多くは，ユーザが必要だと判断した情報から推薦指標を作成している．そのため，推薦指標に合致するが，ユーザに必要ない情報のみを除外することは難しい．

本論文では，ユーザが好むジャンルの中で，ユーザに必要な情報と必要でない情報を抽出する手法を提案する．本手法では，あるジャンルに関して，推薦を受けるユーザと同じジャンルを好むユーザ群を比較し，そのジャンルでのユーザに必要のない情報を予測する．自らと同じジャンルを好むユーザの多くが有益だとしているにも関わらず，推薦を受けるユーザが必要ないとしている情報を抽出し，そのユーザに必要のない情報を表す指標を取り出す．この指標を用いて，ユーザに必要ない情報を除外すれば，ユーザにより精度の高い情報提供ができる．

2. 情報推薦における非有益嗜好の必要性

2.1 情報推薦の適合率と再現率の非両立性

既存の情報推薦手法の多くはユーザの選択行動から指標を作成する．選択行動として，ユーザの web ページの閲覧時間が挙げられる [1]．推薦手法は高い適合率と再現率を保証せねばならないが，適合率と再現率はトレードオフの関係にあり，両方で高い精度を保つことは困難である [2]．

既存手法で web ページを推薦する場合の，適合率と再現率の両立が困難である理由を例を用いて説明する．図 1 では，ユーザが “本田圭佑が日本代表戦で得点を決めた” という web ページを閲覧していたので，ユーザに {“本田圭佑”} という単語を含む web ページを推薦する．推薦指標が {“本田圭佑”} のみなので，推薦指標としては緩いと言える．図 1 の A，B，C の 3 つの web ページがユーザに必要な web ページであった．{“本田圭佑”} という推薦指標に合致するが，D，E の 2 つの web ページはユーザには必要ない web ページだった．このように，ユーザの推薦指標に合致する情報でもユーザに必要な情報とは限らない．

ユーザに必要な情報だけを推薦するために，より厳しい推薦指標を設定することを考える．図 1 の推薦指標 2 のように，{“本田圭佑が得点”}，{“本田圭佑の日本代表は試合”} と設定すれば，A，B のみを推薦できる．しかし，C は推薦結果から除外される．C を推薦できる指標を取り出せる web ページがユーザの閲覧履歴に出現しておらず，推薦指標 2 を用いると，ユーザに必要な C を推薦できない．このように，推薦指標を厳しくすると，ユーザに必要な情報をすべて網羅できず，推薦できない web ページが出てしまう．

†立命館大学情報理工学部
‡立命館大学大学院理工学研究科

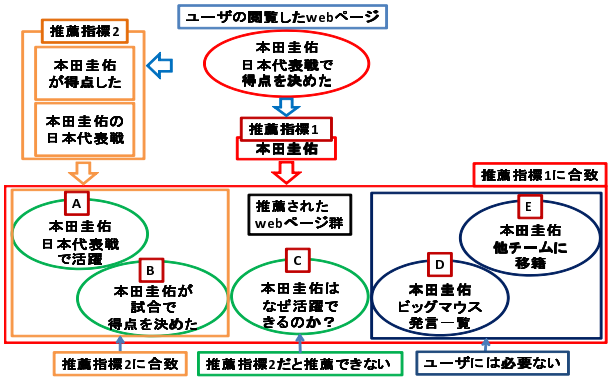

図 1: 適合率と再現率を両立できない推薦例

2.2 有益嗜好と非有益嗜好

推薦指標に合致する情報で，ユーザが必要だと感じる情報には 2 種類考えられる．ひとつは，ユーザの選択行動に出現するものである．図 1 では，ユーザは “本田圭佑が出ている日本代表戦” の情報を必要だと自覚しているので，その情報を有する web ページを閲覧するという選択行動に現れた．もうひとつが，ユーザの選択行動に出現しないものである．ユーザがまだ知らないが，見て初めて必要だと感じるような情報は，ユーザが必要だと自覚できないので，ユーザの選択行動に出現しない．図 1 では，ユーザが “本田圭佑はなぜ活躍できるのか?” という情報を知らず，必要だと自覚できないので，閲覧履歴に現れなかった．どれだけ，ユーザの {“本田圭佑”} という推薦指標を厳しくしても，ユーザの選択行動に出現しない {“本田圭佑”} の web ページは推薦できない．

既存の情報推薦手法の多くは，ユーザが必要だと判断した情報からのみ推薦指標を作成している．しかし，推薦指標に合致するが，ユーザに必要ない情報を予測できないため，推薦指標に合致するが，ユーザに必要ない情報のみを除外することは難しい．そこで，既存の推薦手法に加え，推薦指標に合致するが，ユーザに必要ない情報からも，別の推薦指標を作成する場合を考えよう．図 1 では，D と E がユーザに必要ない情報である．ここで，{“本田圭佑”} に合致する情報は推薦するが，D と E に関する情報は推薦しないという推薦指標を設定する．このような推薦指標ならば，例の中でユーザに必要な A,B,C すべてを推薦できるうえに，ユーザに必要ない D,E を除外できる．このように，ユーザに web ページを推薦するさいに，ユーザに必要のない情報からも指標を取り出せれば，推薦指標に合致する情報の中で，ユーザに必要のない情報のみを除外し，適合率と再現率を両立させることができる．

ユーザの選択行動から，システムが，ユーザが必要だと判断したページを抽出する．さらに，それをもとに，ユーザに必要でないと判断されるページを予測する．あるユーザに必要な web ページを，そのユーザの有益 web ページと定義する．あるユーザの有益 web ページから

抽出した推薦指標に合致はするが、そのユーザに有益でない web ページを非有益 web ページと定義する．ユーザの有益 web ページから取り出す推薦指標を有益嗜好，非有益 web ページから取り出す推薦指標を非有益嗜好と定義する．

3. 類似ユーザと共起語による非有益嗜好抽出

3.1 類似ユーザとの比較による非有益嗜好の抽出

本論文では，情報推薦手法の適合率と再現率を高く保つため，ユーザのブックマークを用いて有益嗜好と非有益嗜好を抽出する手法を提案する．以後，推薦を受けるユーザを被推薦ユーザと呼ぶ．

web ページをブックマークするという行為は，ユーザが後にその web ページを再度訪れる意思表示であると本研究では考える．あるユーザに再び訪れたいと思わせる web ページは，そのユーザに有益な情報を含んでいると言えるので，ユーザがブックマークした web ページは，そのユーザに有益な web ページであると言える．本手法では，ユーザのブックマークを収集するために，ソーシャルブックマーク (SBM) を利用する．SBM とは，ユーザが自身のブックマークを web 上に公開し，不特定多数のユーザとブックマークを共有するサービスで，代表的な SBM の例として，はてなブックマーク [6] が挙げられる．さらに，ユーザは web ページの内容をもとに web ページが有益かどうか判断する．web ページの内容は語の共起の概念によって表現できる [3]．語の共起とは，1 文中である単語 A とある単語 B が同時に出現する概念のことである．ここで，語の共起による 2 つの単語の組み合わせを共起語と定義する．ユーザの推薦指標も共起語の形で抽出できると考えられる．提案手法を適用する前に，ユーザのブックマークをそのユーザの有益 web ページ群として抽出し，各有益 web ページの，共起語群を作成しておく．

図 2 は，ユーザが特定のジャンルに関する web ページの推薦を受けるときに，当該ジャンル下の非有益嗜好を抽出する手法の概要である．被推薦ユーザと同じジャンルに興味を持つ SBM ユーザの有益 web ページを比較することで，被推薦ユーザの有益 web ページ群には明示的に現れていない，当該ジャンル下での非有益 web ページを抽出し，それを用いて非有益嗜好を抽出する．ここで，被推薦ユーザと同じジャンルに興味のある SBM ユーザを類似ユーザと定義する．類似ユーザの多くがブックマークしている web ページは被推薦ユーザもブックマークする可能性が高いと考えられる．類似ユーザの多くがブックマークしているにも関わらず，被推薦ユーザがブックマークしていない web ページは，被推薦ユーザがブックマークしない原因があると考えられるため，被推薦ユーザの非有益 web ページとして抽出する．被推薦ユーザの非有益 web ページから，興味のあるジャンルの web ページであるにも関わらず，被推薦ユーザが有益と判断していない原因を抽出する．

以下，3.2 章，3.3 章，3.4 章で提案手法を説明する．

3.2 有益嗜好の抽出とクラスタの生成

まず被推薦ユーザは，有益 web ページ群から web ページをひとつ選択する．選択された web ページの内容に関連する web ページを推薦することを想定しているためである．例えば，“サッカーの J リーグ” に関する推薦を受けたい場合，被推薦ユーザは “サッカーの J リーグ” の内容が記述してある有益 web ページを選択する．選択された web ページから，ユーザの有益嗜好を抽出する．選択された web ページの内容を表すような共起語が，被推薦ユーザの有益嗜好だと考えられる．

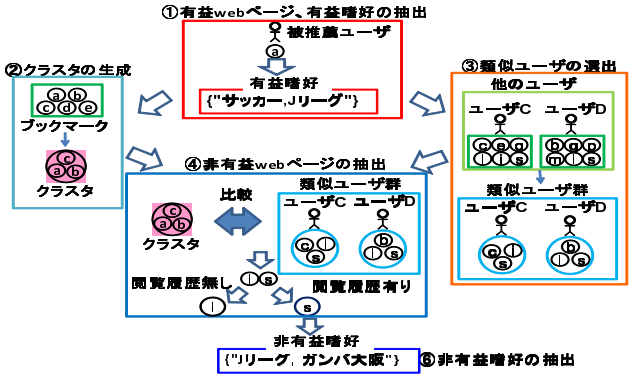

図 2: 提案手法の全体図

有益嗜好の抽出手順を説明する．被推薦ユーザの各有益 web ページに対して作成された共起語にスコアを付与する．例えば，スコアを，タイトルに出現する名詞に 5 点，本文のみに出現する名詞に 1 点と付与する．web ページのタイトルは，web ページの内容を表している場合が多いためである．重み付けに従い，共起語に以下の様にスコアを付ける．

・タイトルに出現する名詞同士が 1 文で共起した場合には 10 点.
・タイトルに出現する名詞と本文のみに出現する名詞が共起した場合には 6 点.
・本文のみに出現する名詞同士が共起した場合には 2 点.

同じ共起語が複数の文に出現した場合，スコアに共起した文の数を掛ける．選択された web ページから抽出された共起語群から，スコアのもっとも大きい共起語ひとつを有益嗜好として抽出する．

次に，被推薦ユーザが選択した web ページのジャンルに対して，どれだけ情報を知っているのかを判定するためにクラスタの作成をする．被推薦ユーザの有益 web ページ群から，有益嗜好に合致する web ページ群を抽出する．この web ページ群をクラスタとする．

3.3 類似ユーザの選出

比較対象になる被推薦ユーザの類似ユーザ群を選出する．被推薦ユーザ以外の各 SBM ユーザの有益 web ページ群から，被推薦ユーザの有益嗜好に合致する web ページを探索する．被推薦ユーザの有益嗜好に合致する web ページをひとつでも有益 web ページとしているユーザを類似ユーザとする．

被推薦ユーザの有益嗜好に合致する web ページは，被推薦ユーザが有益だと判断しているジャンルの web ページと言える．つまり，類似ユーザは被推薦ユーザと同じジャンルを有益だと判断していると言える．

3.4 非有益 web ページと非有益嗜好の抽出

被推薦ユーザの類似ユーザ群を用いて，被推薦ユーザの非有益 web ページを予測し，非有益嗜好を抽出する．多くの類似ユーザが有益 web ページとしている web ページは，被推薦ユーザにも，有益 web ページになる可能性が高い． そこで，多くの類似ユーザが有益 web ページ

としているにも関わらず，被推薦ユーザが有益 web ページとしていない web ページを非有益 web ページの候補とする．非有益 web ページの候補には，被推薦ユーザがまだ閲覧したことのない web ページが含まれることも考えられる．非有益 web ページの候補の中で被推薦ユーザがまだ閲覧していない web ページは，被推薦ユーザが閲覧すれば有益 web ページとする可能性もある．そこで，非有益 web ページの候補から，被推薦ユーザの閲覧履歴がある web ページのみを，各非有益 web ページとして抽出する．閲覧履歴はあらかじめ，被推薦ユーザのブラウザから取得しておいたものを利用する．

非有益 web ページには，クラスタ内の web ページ群には出現しないような共起語を持つと考えられる．全ての非有益 web ページの共起語群とクラスタ内の web ページ群の共起語群を比較し，非有益 web ページのみに出現する共起語群を非有益嗜好として抽出する．

4. 提案手法の有用性の評価

4.1 検証項目

本手法により抽出した非有益嗜好を用いた，推薦手法の有用性を検証するため，今回は本手法における非有益 web ページの条件の妥当性を検証する．以下，非有益 web ページの条件を検証条件と呼ぶことにする．

検証条件は，“多くの類似ユーザが有益だと判断しているにも関わらず，被推薦ユーザが有益だと判断していない web ページ” である．検証条件が妥当ならば，提案手法は，検証条件に合致する web ページから抽出した非有益嗜好を用いて，被推薦ユーザの非有益 web ページを既存の推薦手法の推薦結果から正確に除外できる．検証条件に合致する web ページが以下の 2 点を満たすかを検証し，非有益 web ページを既存の推薦手法の推薦結果から正確に除外できているか検証する．

i 検証条件に合致する web ページから抽出した非有益嗜好ならば，提案手法は，被推薦ユーザに必要のない web ページを多く除外できる．

ii 検証条件に合致する web ページから抽出した非有益嗜好ならば，提案手法は，被推薦ユーザに必要な web ページをほとんど除外しない．

4.2 検証データの収集手順

はてなブックマークユーザの 20 代の男性 5 名を被験者とした．被験者の類似ユーザ候補となる，はてなブックマークの一般利用者 2568 名分のブックマークデータ 89962 個を取得した．本実験ではまず，被験者の有益嗜好を手動で抽出する．有益嗜好を正確に抽出できない場合，被験者に必要な情報を推薦できず，検証項目の検証に支障がでると考えられるためである．被験者に自身の有益 web ページから web ページを 1 つ選択してもらい，その web ページから共起語を作成した．その中で web ページの内容を表していて，興味あると被験者が判断した共起語を 3 から 4 つ選んでもらい，有益嗜好とした．抽出した有益嗜好を用いて，3.2 節の手順に従い，被験者のクラスタの作成と類似ユーザ群の選出をした．次に，各被験者の類似ユーザ群のもつ有益 web ページから，被験者の有益嗜好にひとつでも合致する web ページを被験者に推薦した．最後に，推薦された web ページを，以下の 2 つの項目に関して 4 段階評価のアンケートで評価してもらった．

α “実際に web ページを推薦して欲しいか”
“4：非常に推薦して欲しい”，“3：推薦して欲しい”，“2：推薦して欲しくない”，“1：非常に推薦して欲しくない”．
β “推薦された web ページの内容が有益嗜好を反映しているか”
“4：非常に反映している”，“3：反映している”，“2：あまり反映していない”，“1：非常に反映していない”．

4.3 収集データとその前処理

被験者のアンケートから，web ページの種類を定義する．評価項目 β で 3 または 4 の評価がついている web ページを推薦候補 web ページと定義する．推薦候補 web ページの中で評価項目 α で 1 または 2 の評価がついている web ページを非有益候補 web ページとみなす．推薦候補 web ページの中で評価項目 α で 3 または 4 の評価がついている web ページを有益候補 web ページとみなす．推薦候補 web ページは，本手法を適用しない推薦において被験者の有益嗜好に合致する web ページなので，被験者に推薦される web ページだと想定される．本手法の目的は，推薦される web ページの中で，ユーザの有益嗜好を反映しているにも関わらずユーザに必要のないものを除外することなので，推薦候補 web ページのみを検証対象とする．

4.4 データの検証方法

当該非有益 web ページから生成される非有益嗜好によって，推薦候補 web ページ中の有益候補 web ページをできるだけ除外せず，かつ非有益候補 web ページをできるだけ除外できるかを検証する．ここで，当該非有益 web ページを有益とみなす類似ユーザが多いことは，検証条件により強く適合するとみなせる．より強く検証条件に適合する非有益 web ページから生成される非有益嗜好に，より強く検証項目 i,ii の傾向が見られれば，当条件の妥当性を立証できる．推薦結果のアンケートから抽出された，非有益候補 web ページを非有益 web ページとして想定する．

被験者 5 名の非有益候補 web ページを用いて，被験者の推薦候補 web ページをフィルタリングする．フィルタリングに用いた非有益候補 web ページごとの類似ユーザ数と，フィルタリング結果を比較することで 4.1 節の検証項目 i,ii の 2 点を検証する．

今回，推薦候補 web ページの中で (1) タイトルに出現する名詞同士が共起した場合，(2) タイトルに出現する名詞と本文のみに出現する名詞が 2 文以上で共起した場合，(3) 本文のみに出現する名詞同士が 5 文以上で共起した場合，以上の条件のどれかを満たす共起語が非有益嗜好と合致する web ページを非有益嗜好によってフィルタリングされる web ページとみなし，除外した．

4.5 検証結果と考察

まず検証項目 i を立証するために A:“フィルタリングに使用した非有益候補 web ページの類似ユーザ数” と B:“除外できた非有益候補 web ページ数” を検証する．

A と B の相関を調べると図 3 のような結果になり，0.754 という強い正の相関があることがわかった．ここから，フィルタリングに使用した非有益候補 web ページの類似ユーザ数が増加すれば，多くの非有益候補 web ページを除外できることがわかった．

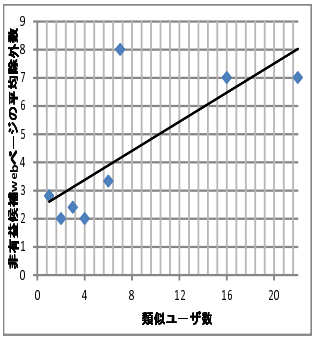

図 3: A と B のグラフ

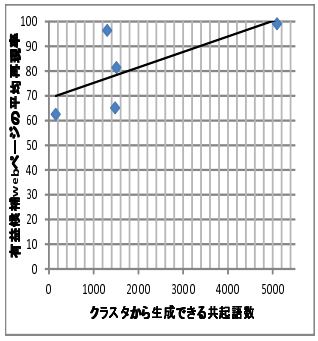

図 4: D と E のグラフ

次に検証項目 ii を立証するために A と C:“除外された有益候補 web ページ数” を検証する．

A と C の相関を調べると，相関係数は 0.09 となった．相関が現れなかった原因として，非有益候補 web ページから抽出した共起語に有益候補 web ページにも頻出する共起語が見られ，有益候補 web ページを除外していた．

有益候補 web ページにも出現する共起語を非有益嗜好としてしまった原因として，クラスタから生成できる共起語数に問題があると考えられる．提案手法は，有益候補 web ページにも出現する共起語を非有益嗜好としないために，非有益候補 web ページの共起語群とクラスタ内の web ページ群の共起語群を比較することで非有益嗜好を作成している．クラスタ内の web ページの共起語数が多ければ，非有益嗜好となる共起語は厳選されるが，少ない場合は有益候補 web ページに出現する共起語も非有益嗜好としてしまう．そこで，被験者ごとに各非有益候補 web ページでフィルタリングしたさいの D:“有益候補 web ページの平均再現率” と E:“各被験者のクラスタから生成できる共起語数” の相関を調べた．平均再現率とは，各非有益候補 web ページでフィルタリングをしたさいに推薦できる有益候補 web ページの再現率の平均値をとったものである．

図 4 は，被験者ごとの各非有益候補 web ページでフィルタリングしたさいの D と E のグラフである．再現率 =(有益候補 web ページ数–フィルタリングで除外された有益候補 web ページ数)/(有益候補 web ページ数) で算出する．図 4 より共起語数が増加すれば，各被験者の平均再現率は上昇することが分かる．被験者ごとの各非有益候補 web ページでフィルタリングしたさいの D と E の相関を調べると，0.688 という正の相関が見られた．つまり，クラスタから生成できる共起語数が増加すれば，有益候補 web ページが除外されにくいということである．今回の実験では，検証項目 ii は立証できなかったが，被験者のクラスタから生成できる共起語数が増加すれば，フィルタリングにより除外される有益候補 web ページ数は減少することが分かった．推薦を受けるジャンルに関する web ページを多く有益 web ページとしていく被推薦ユーザは，クラスタ内の web ページが増加するので，有益候補 web ページが除外されにくくなると思われる．クラスタから生成できる共起語数の多い被験者で再度，検証をすれば検証項目 ii を立証できると考えられる．

以上より，本手法における非有益 web ページの条件は，被推薦ユーザのクラスタから生成できる共起語数が十分な数があれば，効果が期待できる．

5. 関連研究

文献 [4] は，“オンラインニュースサービスにアクセスして最初に提示された記事項目にも関わらず，ユーザが視聴しなかった記事はユーザにとって必要のない記事である” という前提に基づき，前提を満たす記事から指標を取り出して，推薦精度を向上させている．この方法はオンラインニュースのみが対象である．本手法は被推薦ユーザの比較対象となるユーザデータを用意できれば，非有益 web ページの条件を他の推薦手法に応用できるので，本手法は文献 [4] よりも汎用性が高いと言える．

今回は，ユーザが web ページを有益だと判断する基準に，web ページをブックマークしているか否かを利用した．一方で，web ページの閲覧時間から，有益性を予測する研究がされている [1]．web ページを有益だと判断している基準にブックマークだけでなく閲覧時間も用いることができれば，より正確にユーザの有益 web ページを予測できるので，より高精度で非有益 web ページを抽出することが期待できる．

6. おわりに

本論文では，情報推薦手法の適合率と再現率の双方の高い精度を保つために，ユーザのブックマークと協調フィルタリングを用いて有益嗜好と非有益嗜好を抽出する手法を提案した．被推薦ユーザと類似ユーザ群を比較することにより，“多くの類似ユーザが有益だと判断しているにも関わらず，被推薦ユーザが有益だと判断していない web ページ” を非有益 web ページとして予測し，非有益嗜好を抽出することができる．

非有益 web ページの条件の妥当性を検証した結果，クラスタから生成できる共起語数が多い被験者ではあれば，非有益 web ページの条件が妥当であることが分かった．

今回の検証では，被験者数が少なかったため，非有益嗜好を考慮した推薦手法の評価をすることができなかった．今後，被験者数を増やして評価を試みたい．

参考文献

[1] Morita, M. and Shinoda, Y. : information Filtering Based on User Behavior Analysis and Best Match Text Retrieval. Proc. 17th Annual international ACM-SiGiR Conference on Research and Development in information Retrieval, pp.272-281, 1994.

[2] 神嶌敏弘：推薦システムのアルゴリズム (1)．人工知能学会誌，vol.22, no.6, pp.826-837,2007 年

[3] 松尾豊，石塚満: 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム．人工知能学会論文誌．人工知能学会，vol.16, pp.217-223,2002 年 11 月

[4] 大槻一博，服部元，星野春男，松本一則，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/非視聴履歴に基づく嗜好クラスタ管理手法．日本データベース学会 letters．日本データベース学会，pp. 37-40，2007 年．

[5] 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 年．

[6] はてなブックマーク: http://b.hatena.ne.jp/