F-003

意外性ある推薦のための非有益性に基づく情報フィルタリング
Useless Information Filtering for Serendipity

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

1. はじめに

WWW の普及により，個人が手軽に情報を取得できるようになった．一方で，取得できる情報の量があまりにも膨大すぎるために，ユーザ自身に必要な情報を適切に取捨選択することが困難になっている．そのため，ユーザの興味に合った情報を推薦する web ページの推薦システムの重要性が増している．

既存の推薦システムによる情報推薦は，ユーザがこれまでに有益だと感じた情報に合致した情報が推薦される．そのため，既存の推薦システムによる情報推薦によって推薦される情報は，ユーザにとって既知の内容であることも多い．

推薦システムにとってユーザがこれまでに有益だと感じた情報に合致する情報を推薦することは重要である．同様に，ユーザが有益だと気づいていないが，閲覧して初めて有益だと感じる意外性のある情報を推薦することも重要である．

あるユーザにとって意外性のある情報は，情報全体からそのユーザにとって有益でない情報を除外することで，取得できる可能性が高くなると考えられる．推薦を受けるユーザにとって，有益でない情報を除外するための指標が必要である．本論文では，意外性のある情報を推薦するために，ユーザにとって有益でない情報を抽出する手法を提案する．

2. 非有益嗜好と語の共起

あるユーザにとって有益でない情報を表す指標を，そのユーザの非有益嗜好と定義する．ユーザは web ページの内容をもとに web ページを有益でないと判断する．そのため，非有益嗜好はユーザが有益でないと判断した web ページの内容から求められると考えられる．ユーザに有益な web ページを有益 web ページ，有益でない web ページを非有益 web ページと定義する．

web ページの特徴は，語の共起によって表現できると考えられる [2]．語の共起とは，一文中にある単語 A とある単語 B が同時に出現することである．推薦を受けるユーザの非有益嗜好は，推薦を受けるユーザにとって有益でない情報を特徴づける 2 つの単語の組み合わせによって表現できると考えられる．よって，推薦を受けるユーザの有益でない情報において頻出する共起語の集合は，そのユーザの非有益嗜好を表していると言える．

3. 共起語を用いた非有益嗜好の抽出

3.1 抽出の流れ

本論文では，意外性ある情報を推薦するために，推薦を受けるユーザの非有益嗜好を共起語の集合として抽出する手法を提案する．本手法は，ソーシャルブックマークを利用して，被推薦ユーザの非有益嗜好を導出する．

†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

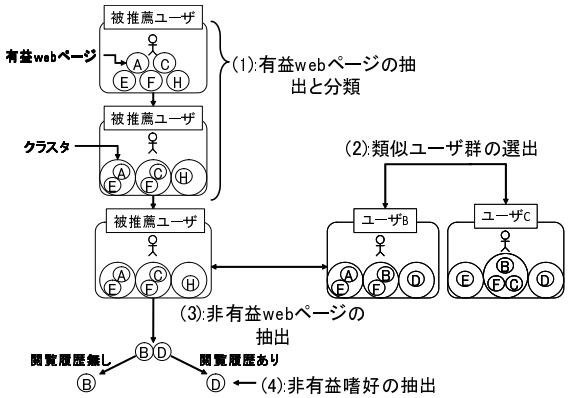

図 1: 手法の全体像

手法の概要は図 1 の通りである．被推薦ユーザは 1 つ以上のジャンルに対して，そのジャンルに関する情報を有益と考え，web ページをブックマークしたり，web ページを閲覧する．この時，被推薦ユーザと同じジャンルの情報を有益と考えているユーザ群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，被推薦ユーザが有益でないと判断する内容が含まれていると考えられる．被推薦ユーザが有益と判断するあるジャンルと同じジャンルの情報を有益と判断するユーザを，類似ユーザと定義する．

3.2 有益 web ページ群の抽出と分類

ユーザがブックマークしている web ページは，そのユーザにとって有益であることを明示的に示されていると考えられる．文献 [1] において，ネットニュースの記事における閲覧時間の長さ，その記事の有用さの度合いに正の相関があることを示す結果が報告されている．よって，以下の 2 つのうち少なくとも片方を満たす web ページを有益 web ページとして抽出する．

(1) ユーザがブックマークしている web ページ．

(2) ユーザが T 秒以上かけて閲覧している web ページ．ただし T は閾値である．

各 web ページから抽出された共起語群を用いて階層的にクラスタリングをする．各有益 web ページに対して形態素解析をして，一文ごとに共起語群を抽出する．有益 web ページごとに抽出した共起語群と，その出現数を比較する．この時，以下の条件を両方とも満たす共起語を見つけ，その共起語を持つか否かで有益 web ページ群を分割する．

(1) 有益 web ページの内容を特徴づける共起語群の中で，もっとも多くの有益 web ページで共通する．

(2)(1) の共起語が有益 web ページ群の中で頻出する．

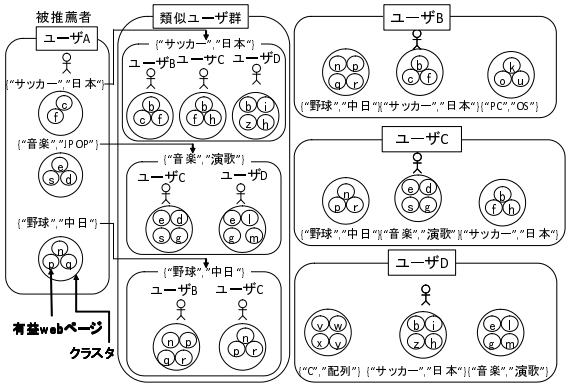

図 2: 類似ユーザ群の抽出

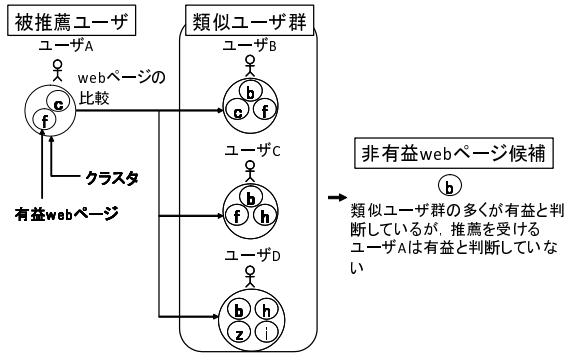

図 3: 類似ユーザ群との比較

この作業を，まだクラスタリングに使用されていない共起語について，有益 web ページ群が分割できなくなるまで再帰的に実施する．分類に使用された共起語をクラスタのラベルと定義する．

3.3 類似ユーザ群の抽出

類似ユーザ群とは被推薦ユーザと同じラベルのクラスタを持つ類似ユーザの集合である．類似ユーザは，被推薦ユーザの持つクラスタごとに選出する．これによって，図 2 のように被推薦ユーザの持つ各クラスタについて，複数の類似ユーザが抽出される．

3.4 非有益嗜好の抽出

有益 web ページ群の各クラスタについて，多くの類似ユーザが有益と判断している web ページは被推薦ユーザにとっても有益 web ページとなることが多い．多くの類似ユーザで有益 web ページとして抽出されているにも関わらず，被推薦ユーザでは有益 web ページとして抽出されていない web ページには，被推薦ユーザが有益でないと判断する内容が含まれている可能性がある．ここで，被推薦ユーザと類似ユーザ群のユーザが持つクラスタの中で，同じラベルのクラスタ同士を比較する．クラスタの比較によって，類似ユーザ群内で多くのユーザの有益 web ページとして抽出されるが，被推薦ユーザの有益 web ページとして抽出されていない web ページを発見する．ここで，発見された web ページを非有益 web ページの候補とする．

図 3 の例では，類似ユーザ群の多くがアイテム b を有益と判断しているにも関わらず，ユーザ A は有益と判断していないので，アイテム b はユーザ A の非有益 web ページの候補となる．

非有益 web ページの各候補に対し，被推薦ユーザが本当に非有益と考えてブックマークしていないのか否かを区別する必要がある．そこで，被推薦ユーザの閲覧履歴を参照し，被推薦ユーザが抽出された各非有益 web ページの候補を閲覧しているかどうか確認する．被推薦ユーザがまだ閲覧していない web ページを非有益 web ページ候補から除外して，被推薦ユーザが過去に閲覧している web ページ群を非有益 web ページ群とする．

非有益嗜好を表す指標は，共起語を用いて表現できると考えられる．非有益嗜好を表す共起語は，非有益 web ページ群内において，頻繁に出現する共起語であると考えられる．しかし，共起語の中には，どの web ページにも頻繁に出現するような共起語も存在する．非有益嗜好を抽出するために，被推薦ユーザの有益 web ページ群において出現する共起語と，被推薦ユーザの非有益 web ページ群の共起語を比較する．被推薦ユーザの有益 web ページ群においてあまり出現せず，非有益 web ページ群において頻繁に出現する共起語を発見し，非有益嗜好として抽出する．例えば，被推薦ユーザの非有益 web ページ群から抽出した共起語に限り {“J リーグ”，“ガンバ大阪”}，{“ガンバ大阪”，“選手”} という共起語が出現している，非有益嗜好を表す共起語としてこの 2 つの共起語を抽出する．

4. 関連研究

文献 [3] では，推薦精度を向上するために，ユーザが視聴しなかった記事の履歴である，非視聴履歴を利用した手法を提案している．ユーザが有益でないと判断した情報を指標として用いる点では，本研究と類似しているが，文献 [3] はネットニュースの購読者という限定的な状況でのみ指標を取り出すことが可能になる．その点で本研究はブックマークと閲覧履歴から指標を抽出するので，文献 [3] より汎用性が高いと言える．

5. おわりに

本論文では, 意外性のある情報を推薦するために, 非有益嗜好を抽出する手法を提案した．今後は, 本手法の有用性の検証する予定である．

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] 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

[3] 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