社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－DBS－143（20）
2007／7／2

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

高須賀清隆† 丸山 一貴†† 寺田 実†††

† 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 〒182-0026 東京都調布市調布ヶ丘1-5-1
†† 電気通信大学情報基盤センター 〒182-0026 東京都調布市調布ヶ丘1-5-1
††† 電気通信大学電気通信学部情報通信工学科 〒182-0026 東京都調布市調布ヶ丘1-5-1
E-mail: †{takasuka,terada}@ice.uec.ac.jp, ††kazutaka@acm.org

あらまし 近年，情報爆発に伴い，多量な情報の中から情報を取り出すシステムの研究が盛んに行われている．その中でWebページ推薦の分野ではユーザのブックマークを利用したシステムが多く，ブックマークされないようなWebページを推薦できないという問題があった．そこで本研究では，ユーザたちの閲覧履歴そのものを利用することで全Webページを対象に推薦可能とするシステムを構築し，その評価を行う．

キーワード Webページ推薦，閲覧履歴，協調フィルタリング，単語抽出による自動評価

Web page recommendation by URL-based collaborative filtering and its evaluation

Kiyotaka TAKASUKA†, Kazutaka MARUYAMA††, and Minoru TERADA†††

† Department of Information System Fundamentals, The University of Electro-Communications Chofugaoka 1-5-1, Chofu, Tokyo, 182-0026 Japan
†† Information Technology Center, The University of Electro-Communications Chofugaoka 1-5-1, Chofu, Tokyo, 182-0026 Japan
††† Department of Information and Communication Engineering, The University of Electro-Communications Chofugaoka 1-5-1, Chofu, Tokyo, 182-0026 Japan
E-mail: †{takasuka,terada}@ice.uec.ac.jp, ††kazutaka@acm.org

Abstract Because the number of web pages becomes very huge, and still increasing, many people have difficulty to reach pages they want. Although social bookmarking and search engines are helpful, users still have to find pages by themselves. Our goal is to recommend web pages which are supposed to be interesting for a user, without any extra actions of users. We developed a recommendation system that works based on URLs and the users. Our system has four features: (1) collaborative filtering based on URL only, (2) similarity between users using TF-IDF, (3) use of the real activity in our university, (4) and automatic evaluation using word extraction.

Key words Web page recommendation, web browsing history, collaborative filtering, automatic evaluation using word extraction

1. 背 景

近年インターネットは一般にも大きく普及し，ネットワークインフラも整い，テキストのみならず動画や画像なども含めてWeb上の情報量は膨大なものとなってきている．現状ではこの膨大な量の情報をユーザ個人個人が効果的に利用できているとは言いがたい．多くのユーザは日々特定のWebサイトを巡回するだけであったり，ちょっとした調べ物をする際に使用するのみに留まってしまっている．しかし，そのユーザが興味を抱くであろう情報を持つページが存在する可能性は高く，その存在を見過ごしているような状況にある．

その状況を改善するために，検索エンジンやソーシャルブックマークなどがあるが，それらとは違う尺度でWebページを探索する手段としてWebページ推薦システムが研究されている[1][2][3]．

従来の研究の多くはユーザのブックマークを利用したものであり，推薦対象はブックマークされたページに制限されたものだった．ブックマークに限定せずにWeb上の全てのページを

対象とする研究もあるが，専用のブラウザを利用する必要があるなどユーザに対する負荷が大きかった．

2. 目　　的

本研究の目的は，ブラウザの拡張機能を用いてユーザの Web 閲覧履歴を自動的に収集し，閲覧行動が類似するユーザの履歴から Web ページを推薦するシステムを構築し，評価することである．ブックマークを利用するシステムとは異なり，ブラウザの拡張機能を用いて履歴を収集することで，ユーザはシステム利用時にほぼ負荷を受けることなく，普段通りに Web を閲覧することが可能である．ブックマークを利用するものよりも，より負荷が軽く，よりタイムリーな話題を扱った新しい Web ページを推薦したい．

3. 関 連 研 究

3.1 技術的課題

Web が持つ特徴により Web ページ推薦には課題が二つある．

（1） 推薦システムを用いているので有名な Amazon.com で扱っているアイテムの量は数百万ほどである [4] が，Web 上に存在する Web ページの総量は数百億とも言われ，これまで開発されている推薦システムが扱うアイテム数に比べて膨大な量である．

（2） アイテムである Web ページは無数の Web サーバ上に分散配置されており，アクセスの検出は Web ページが存在する Web サーバ以外では難しい．

3.2 先 行 研 究

3.2.1 特定サイト限定型

利用を特定のサーバに限定する方式である．Li らによる [2] があげられる．Web サーバのアクセスログを入力として，コンテンツの内容や構造を加味してユーザの行動パターンを分析し，Web ページ推薦に利用している．しかし，Web サーバのアクセスログを入力として使用するため，特定の Web サイトに利用が限定されてしまう．

特定サイトに限定することで技術課題 1，2 を共に克服していると言える．

3.2.2 Web 全体型

特定の Web サイトに利用を限定せずに Web 全体を推薦対象とできる方式である．Zhu らによる [1] が挙げられる．このシステムは Web 全体を推薦対象とすることができるという点において本研究と類似している．しかし，システムを利用するには専用のブラウザなどを使用する必要がある．また，Web ページを評価するために操作が必要である．[1] は推薦の生成には Google のような検索エンジンを利用している．ユーザの閲覧履歴を元にユーザが現在閲覧しているページの代表的な単語を抽出し，決定木を用いて関連単語を類推する．後は関連単語からクエリーを生成して検索エンジンに送る．この例では専用ブラウザを用いることで技術課題 2 を克服しており，検索エンジンを利用することで技術課題 1 を克服している．

本研究では Web ページ閲覧を暗黙的な評価として扱うためユーザに負荷を与えないという点で優位であると言える．

3.2.3 ブックマーク利用型

ユーザの興味の表れとしてブックマークをユーザプロファイルを表すために利用する方式である [5] [6]．ブックマークを利用することでユーザプロファイルの作成を容易にする．しかし，ブックマークを用いる手法は，ブックマークする作業が必要である．そのため，より細かいユーザプロファイルを作るには Web 閲覧の際に頻繁にブックマークしなくてはならずユーザの負担が大きい．

Web 全体ではなくて，ブックマークを使っていることで Web ページの量を圧縮し技術課題 1 を克服している．また，ブックマークの情報さえあれば推薦できるため，Web ページへの実際のアクセスをシステム側で知る必要はなく，技術課題 2 とは無縁である．

4. 本研究の特徴

本研究の主な特徴を以下にまとめる．

（1） URL のみによる推薦生成．

（2） 有名サイトの除外．

（3） TF-IDF 法 [7] を応用したユーザ間類似度．

（4） 実データを用いた検証．

（5） 単語抽出を用いた自動評価手法．

本研究では，収集された閲覧履歴を元に推薦を決定する．その際に，システムでは Web ページの内容を解析するなどは行わず，URL の列としての閲覧履歴だけを元に推薦を生成している．この点が本研究の大きな特徴である．

例えば検索エンジンのトップページなどのように，履歴に存在していなかったとしても推薦される必要のないほどに有名な Web ページなどが存在する．そういった Web ページは推薦する必要がないため，推薦からは除外するようにしている．

推薦は協調フィルタリングを用いて生成する．推薦を受けるユーザと類似したユーザを見つける際に利用するユーザ間の類似度に，頻度と一般性を考慮する TF-IDF 法の考え方を用いている．

検証実験は電気通信大学の対外接続部に設置したスニッファにより収集された HTTP トラフィックデータを用いて行っている．実際のシステムではブラウザの拡張機能から閲覧履歴を収集することを想定している．しかし，閲覧履歴収集用のブラウザ拡張を大人数に普及するのは難しいため実験用の措置としてスニッファの出力を利用している．

Web ページからの単語抽出を行い，推薦から抽出された単語群と履歴から抽出された単語群の比較を行うことで評価する自動評価を行う．

本研究では，第 3.1 節で記述した技術課題 1 に，推薦対象を一度でも閲覧されたページとすることにより対応している．これにより，ブックマークを利用するタイプのシステムよりもより多くの Web ページを対象とすることが可能である．技術課題 2 については，閲覧履歴をブラウザの拡張から送信してもらうことで解決したい．

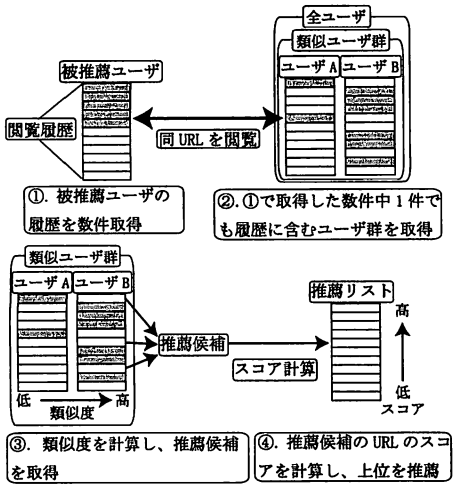

図 1　アルゴリズムの概略

5. 推薦アルゴリズムの概要

推薦を生成するアルゴリズムをステップに分けて説明する．

(ステップ 1)　推薦対象ユーザの履歴から最新の URL(以下プローブ) を取得．

(ステップ 2)　ステップ 1 で取得したプローブ中 1 件でも履歴に含むユーザ群 (以下類似ユーザ群) を取得．ユーザ群が 20 人に満たない場合はプローブを 1 件ずつ増やした．

(ステップ 3)　ステップ 2 で取得した類似ユーザ群と推薦対象ユーザの類似度を計算し，推薦候補を取得．

(ステップ 4)　各推薦候補のページスコアを計算し推薦を決定．

アルゴリズムの概略を図 1 に示す．

以下では第 5.1 節で本研究で用いているユーザの興味を表すユーザプロファイルについて説明する．第 5.2 節で類似ユーザ群について説明し，第 5.3 節で興味の近さを表すユーザ間類似度を定義し，第 5.4 節で推薦候補の決定法について説明し，第 5.5 節でページのスコアについて説明する．

5.1　ユーザプロファイル

本研究ではユーザプロファイルとしてユーザのページの閲覧履歴である URL のみを使用する．

URL のみを用いるのには利点がある．閲覧したページの内容を解析してユーザプロファイルを作成する場合，ページのテキストを解析するため，テキストを含まないページは推薦にすることができないという問題がある．本推薦アルゴリズムではページの内容を利用しないため，あらゆるページを推薦対象とすることができる．また，日本語のような非分かち書き言語や未知の単語にも対応できるという利点もある．

閲覧履歴そのものをユーザプロファイルとして利用することには以下のような利点がある．

（1）　変化の激しい短期的な興味さえも全て記録されている閲覧履歴を用いることで興味の変化にも対応できる．

（2）　Web 閲覧を行うだけで構築されるためユーザに特別な入力を必要とさせない．

利点 1 について説明する．ユーザの興味は頻繁に変化し得るものであり，普段はまったく興味がないけれどもちょっとした興味でページを閲覧する場合もある．明示的にユーザプロファイルを作成する場合は，このような細かい興味の変化に柔軟に対応できないが，閲覧履歴をそのまま用いることで暗黙的にユーザプロファイルを作成できる本研究では，履歴の最新の数件を取り出してプローブとするなどの操作によって現在のユーザの興味を取り出すことができる．また，長期間の履歴を参照することでそのユーザの長期的な興味を知ることもできる．利点 2 に関して説明する．Web ページの内容に興味があったか否かなどの評価をユーザは一切する必要がないため，ユーザの Web 閲覧を妨げないということである．

ステップ 1 で取得するプローブは，ユーザの短期的な興味を推薦に反映させる目的で取得している．

5.2　類似ユーザ群

システムはステップ 1 でプローブを取得し，ステップ 2 でプローブを 1 件でも閲覧したことのあるユーザを全てリストアップする．本システムではこのユーザ群を類似ユーザ群と定義し，推薦対象ユーザと興味が重なったユーザたちであると考える．

5.3　ユーザ間類似度

ステップ 3 で計算するのがユーザ間類似度である．TF-IDF 法と同様に，頻度と一般性を考慮して，以下の二点を元に計算している．

- ユーザ間で共通して閲覧された URL 数．
- 共通して閲覧された各 URL を閲覧したユニークユーザ数．

第 1 項が TF-IDF 法における TF に相当し，この数が多いほどユーザ間類似度は大きくなる．第 2 項の逆数を取ったものが TF-IDF 法の IDF に相当し，この値で各 URL を重み付けする．

ユーザ a の閲覧履歴の集合を $hist(a)$，ユーザ b の閲覧履歴の集合を $hist(b)$ とする．ページ x を閲覧したユニークユーザ数を $uusr(x)$ として以下の式でユーザ a とユーザ b の類似度 $similarity(a,b)$ を計算する．

$$similarity(a,b) = \sum_{x \in hist(a) \cap hist(b)} \frac{1}{uusr(x)}$$

5.4　推薦候補の決定

ステップ 3 で類似ユーザ群の各ユーザと推薦対象ユーザの類似度を計算した後に推薦候補を決める．推薦候補は類似ユーザ群中で最高類似度のユーザの履歴を元に作成される．最高類似度のユーザの履歴から推薦対象ユーザの履歴に含まれる URL を取り除き推薦対象ユーザが閲覧したことがない URL のみにする．その URL 群を推薦候補とする．

5.5　ページスコア

まず類似ユーザ群の数と同数の一般ユーザ群を取得する．一般ユーザ群は類似ユーザ群に含まれないユーザをランダムに取得した集合である．2 つのユーザ群は図 2 に示すように 1 対 1 に対応付ける．各ユーザは投票値を持っており，その投票値は以下のように決定される．

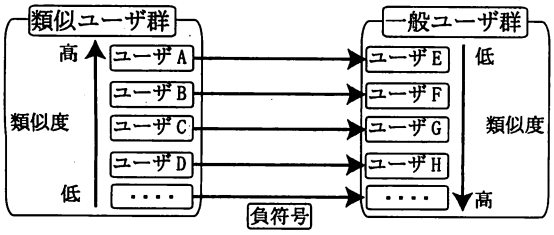

図 2　一般ユーザ群と類似ユーザ群の対応表

(類似ユーザ群)　推薦対象ユーザとの類似度そのもの.
(類似ユーザ群中最高類似ユーザ)　1.0.
(一般ユーザ群)　対応付けられた類似ユーザ群のユーザの類似度の負の値.
(計算対象の URL を閲覧していないユーザ)　0.

類似ユーザ群中の最高類似ユーザの投票値を 1.0 にするのは，このユーザの影響が推薦候補を決める段階で大きく入るので，この段階では極力影響を下げるためである.

スコアを計算するページ x，類似ユーザ群と一般ユーザ群でページ x を閲覧したユーザの集合 B，ユーザ b の投票値 $votepoint(b)$ とするとスコアは以下のように計算される.

$$score(x) = \frac{\sum_{b \in B} votepoint(b)}{uusr(x)}$$

こうして計算されたページスコアのうち上位 30 件が推薦としてユーザに提供される.

5.6 計 算 量

本システムにおける計算量を示す.

ユーザ数 N，最長閲覧履歴 M とすると類似ユーザ群の類似度を計算するのに O(NM) かかる．また，推薦候補のスコアを算出するのに掛け算方式，投票方式ともに O(NM) かかる．よって推薦を生成するのにかかる計算量は O(NM) となる.

6. 実　　験

6.1 主観評価に基づく検証

電気通信大学の対外接続部に設置したスニッファの出力データ 1 日分の提供を受け，そのデータと 9 人の被験者の 1 日分の履歴を用いて各ユーザに対して推薦を生成した．また，9 人の被験者の履歴はプロキシを用いて取得した.

推薦された URL それぞれについて以下の項目を 1〜4 の 4 段階で評価してもらった．数値が高いほど評価が高い.

- 自分の興味に沿う内容である.
- 面白かった.

「自分の興味に沿う内容で面白かった」という評価がされたケースと，「自分の興味には沿わない内容だが面白かった」という評価がされたケースを推薦成功と定義した．スニッファで取得されたデータを実験に使用しているため，ユーザには直接関係のないソースコードなどの URL が推薦に含まれてしまう．推薦 30 件中でこれらの URL が含まれる割合を不要 URL 率とし，これらの URL を除いた推薦全体における成功の割合を成功率とした.

表 1　主観評価結果 1(1:悪い，4:良い)

興味に沿う	面白かった	また推薦を見たい	全体評価
2.13	2.2	2.78	2.33

表 2　主観評価結果 2[%]

不要 URL 率	成功率
0.19	0.35

また，30 件の推薦全体として以下の項目について 4 段階で評価してもらった.

- また推薦を見たい.
- 全体としての評価.

表 1，表 2 に結果を示す.

各評価は 1〜4 なので，中央値は 2.5 になる．それを踏まえて結果を見ると，また推薦を見たいかという質問についてはやや良い評価を受けており，全体評価についてはやや悪い方に傾いている．被験者の意見で，同じサイトの似たような Web ページばかりが推薦された，同じ種類のページが多いという意見を受けたが，そのような問題が原因で全体評価が低下していると考えられる．このことから，推薦 30 件の中に同一サイトからの Web ページを複数含むことは避けて，可能な限り多種類の内容を扱う Web ページを揃えることが全体評価の向上に繋がるのではないかと推測している.

不要 URL 率は約 2 割なので，推薦 30 件中で平均して 6 件ほど不要な URL が含まれる計算になる．不要 URL はスニッファからの出力データを用いた実験固有の問題であり，本システムが想定しているブラウザから履歴を取得する方式ではこの問題は起こり得ない．そのため本質的には問題ではないが，今後実験を続けることを考えれば，より精確にシステムの性能を測るために不要 URL を排除することは重要である.

不要 URL を推薦 30 件から除いた有効推薦中における成功した推薦の割合が成功率であり，今後はこの値と全体評価を向上させることが主な目的となる.

6.2 単語抽出による評価

評価に形態素解析と TF-IDF を用いた単語抽出を用いている．形態素解析エンジンはオープンソース形態素解析エンジン MeCab [8] を利用しており，ユーザの履歴一件一件をダウンロードして単語抽出を行う．そうして推薦や履歴から抽出された単語を集計した単語群をそれぞれ推薦キーワード集合，ユーザキーワード集合とする．そして，各キーワード集合から各被験者への推薦それぞれについて適合率と再現率を計算した.

適合率と再現率の説明の前に，ユーザキーワード集合と推薦キーワード集合と呼んでいる二つのキーワード集合を説明する.
(ユーザキーワード集合)　ユーザの興味を表すキーワード集合．推薦対象ユーザの履歴一つ一つに対し単語抽出を行い，1 ページにつき TF-IDF 値の上位 10 件を取得する．この単語群をユーザキーワード集合とする．履歴が 100 件あれば最大で 1000 単語がユーザキーワード集合となる.
(推薦キーワード集合)　推薦群の内容を表すキーワード集合.

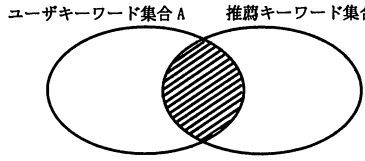

図 3　適合率と再現率の関係

表 3　単語抽出による評価値の平均 [%]

適合率	再現率
17.74	6.58

表 4　推薦生成過程の各値の平均

類似ユーザ群数	最高類似度	最高ページスコア
132.67	4.17	1.57

最終的に推薦された 30 件の URL から，ユーザキーワード集合の場合と同様に，1 ページにつき 10 個ずつキーワードを抽出する．その単語群を推薦キーワード集合とする．30 件の Web ページから 10 個ずつなので最大で 300 単語が推薦キーワード集合となる．

次にユーザキーワード集合と推薦キーワード集合を作成した後に，適合率と再現率をそれぞれ算出する．本論文での適合率と再現率は以下のように定義する．また，図 3 に両者の関係を示す．

(適合率)　ユーザキーワード集合を推薦対象ユーザが望む単語，推薦キーワード集合を結果と捉え，推薦キーワード集合中でユーザキーワード集合に含まれる単語の割合である．適合率はユーザキーワード集合に含まれない単語をノイズとみなすと推薦の精度と考えられる．ユーザキーワード集合に含まれない単語が抽出されるページ，すなわち普段閲覧しているページに似ていないページが推薦中に少ないと上昇する．

(再現率)　ユーザキーワード集合を結果として出るべきもの，推薦キーワード集合を結果と捉え，ユーザキーワード集合中で実際に推薦キーワード集合に含まれた単語の割合である．再現率は推薦によるユーザの興味のカバー率だと考えられる．ユーザキーワード集合に含まれる単語を抽出されるページ，すなわち普段閲覧しているページに似たページが推薦に多く含まれていると上昇する．

適合率，再現率の平均値を表 3 に示す．

また，推薦生成の過程で計算する類似ユーザ群の数，最高類似ユーザの類似度，最高ページスコアの値の平均値を表 4 に示す．

次節にて表 1，表 2，表 3，表 4 の結果の相関関係を求めることで，本手法がどの程度ユーザの感覚を反映できるのかを検証する．

6.3 考　　察

測定した各値間の相関係数を算出した．表 5 に示す．

表 5 を見ると，「また見たいと適合率」，「最高類似度と成功率」の二組について特に高い相関が見られた．

適合率とまた見たいという主観評価に高い相関があるということは，推薦に自分が興味ある単語が複数含まれると推薦をより欲しいと感じることを意味する．これは，自分が興味ある単語が画面に表示されているため，もっと自分の興味ある内容が推薦に出てくるかもしれないという期待感が増すためではないかと考えられる．また，興味ある単語が少ない場合は自分の興味が反映されていないと判断してしまい，推薦への期待感が薄れているのだと推測できる．以上より，推薦には興味ある単語をタイトルなどに多数含むことが，ユーザにシステムを使い続けてもらうという点で重要な要素になると思われる．「最高類似度と適合率」にもやや相関が見られるため，最高類似度の向上が適合率の向上に繋がり，適合率の向上がユーザの期待感を煽ることに繋がるのではないかと考えられる．

最高類似度と成功率の相関が高いということは，類似ユーザ群中での類似度が高ければ高いほど，ユーザの主観として推薦の精度が高くなるということであり，推薦 30 件中にユーザが面白いと感じる Web の割合が大きくなるということである．そもそも推薦候補は最高類似ユーザの履歴を元に生成しているため，このユーザの類似度が高くなると推薦候補に推薦対象ユーザの興味に近い Web ページが含まれる可能性が高くなる．そのため推薦対象ユーザにとって評価の高い Web ページを多く推薦するには類似ユーザ群中に高い類似度を持つユーザを見つけることが重要なのだと考えられる．また，このデータに強い相関が現れていることは，本研究の特徴である TF-IDF 法を応用したユーザ間類似度の計算法が，ユーザ間の類似度を的確に表せている結果だと捉えることができる．

しかし，「成功率と全体評価」の相関係数は 0.26 という結果となっており，推薦成功率の高さが推薦全体としての評価に直接強く結びついていないことが興味深い．全体評価と強い相関があるデータはまだ取れていないが，全体評価の向上には少数でもよいので，ユーザにとって極めて良い印象を与える Web ページが推薦されることが重要なのではないかと推測している．

「類似ユーザ群数とまた見たい」，「再現率と成功率」にやや負の相関が見られた．

「類似ユーザ群数とまた見たい」の組み合わせに負の相関が見られるのは次のような理由が考えられる．類似ユーザ群数が多い場合は有名サイトがプローブに入る場合が多く，ユーザと閲覧行動が類似していないユーザも多数類似ユーザ群に入り込んでいることが予想される．そのため，推薦に推薦対象ユーザの興味ある内容の Web ページが含まれなかった可能性が高い．

「再現率と成功率」についてはこれといった理由が思い当たらず今後詳しく調査する必要がある．

7. 結　　論

本論文では，閲覧履歴を自動的に収集し Web ページを推薦するシステムの推薦生成部分の実装を行い，評価を行った．

ユーザが興味ある分野の単語が推薦に増えると，ユーザの期待感が大きくなることから，本論文で提案する自動評価手法による評価値である適合率がユーザの期待感を煽れているかどうかの指標となれることが示すことができたと言える．また，最

表 5 各値間の相関係数

	類似ユーザ群数	最高類似度	最高ページスコア	再現率	適合率	また見たい	全体評価	成功率
類似ユーザ群数		0.09	-0.15	-0.09	-0.45	-0.62	0.1	-0.03
最高類似度			0.07	-0.35	0.56	0.15	0.17	0.88
最高スコア				0.26	0.47	0.45	0.36	-0.22
再現率					-0.01	-0.23	-0.39	-0.57
適合率						0.71	-0.11	-0.39
また見たい							-0.07	0.15
全体評価								0.26
成功率								

高類似度の大きさが成功率向上の重要な要素となっていることが相関関係から明らかになったことで，より高い類似度を持つユーザを見つけることで推薦成功を増やせると考えられる．

最高類似度と適合率の相関もあるため，最高類似度はまた見たいというユーザによる主観評価にも繋がりうる．そのため，最高類似度は「また見たい」,「全体評価」の両方の主観評価に対して強く影響を与えていることが認められた．

今後，最高類似度の向上を狙うことでユーザの主観評価を向上させることができると考えている．

8. 今後の課題

本研究の課題がいくつか明らかになった．それらを以下に示し，それぞれについて現状考えられる改善案を示す．

(類似度計算) 現在のシステムでは推薦を生成するたびに類似度を計算しているため，閲覧履歴が非常に長いユーザや類似ユーザ群が非常に大きくなってしまった場合などに，推薦の生成に時間がかかってしまうという問題がある．これは実際にシステムを運用する場合などには定期的にユーザ間類似度を計算しておくなどの措置により解消できる．

(類似ユーザ群の補正) 類似ユーザ群の最高類似度が推薦の成功に影響を与えている．類似度の分布が低めに偏っていたりした場合などに推薦の成功率が低下してしまう可能性が高い．ゆえに推薦の成功率が上がるように類似ユーザ群の類似度の偏りを補正するようにユーザを追加することで推薦の評価を向上させることができると思われる．

(新たなパラメータの導入) 類似度の計算やページスコアの計算にはまだ課題も多いため，今後は計算法の改善も必要である．現在は URL しか参照していないので，今後は履歴の順番や，閲覧された時間，どの URL から辿ってきたかなどの情報を計算に反映することを考えている．現状では履歴の新旧を計算には考慮していないため，過去の閲覧の影響がいつまでも残ってしまう問題が確認できている．これは時間情報などを導入していくことでうまく反映させられていないユーザの短期的な興味を推薦に反映させられると考えている．

(不要な URL の除去) 実験には電気通信大学にてスニッファで取得されたデータを用いている．そのため広告やアクセス解析用の 1 ピクセル画像などのページに付随して GET されるデータが数多く推薦として出てきてしまう問題がある．また，それらのデータの影響で類似度が大きくなってしまうなどの問題もある．これらの問題はブラウザから取得したユーザデータを用いることで改善できる問題である [9].

(単語抽出精度の向上) 各 Web ページから代表的な単語を抽出するが，例えば掲示板のスレッドから抽出された単語が書き込んだユーザのデフォルトの名前だったり,「投稿」や「表示」といった一般性の高い単語だったりとうまく抽出できていないと思われるところが多々見受けられる．ユーザキーワード集合と推薦キーワード集合にも影響がある点なので正しく評価するためにも単語抽出精度の向上は必要である．

(全体評価向上) 今回の実験では全体評価が向上するための具体的な要素を知ることはできなかった．今後は全体評価向上の鍵となるのはどういう要素なのかを実験を通じて明らかにする必要がある．

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] R.Greiner, T.Zhu, G.Haubl, K.Jewell. A Trustable Recommender System for Web Content. *Beyond Personalization 2005*, pp. 83–88, Jan 09 2005.

[2] Jia Li and Osmar R. Zaïane. Combining Usage, Content, and Structure Data to Improve Web Site Recommendation. In *EC-Web*, pp. 305–315, 2004.

[3] 佐世保圭介, 波多野賢治, 宮崎純, 吉川正俊, 植村俊亮. ブックマークの階層構造情報を組み込んだ協調フィルタリングによる web ページの推薦手法,. 第 15 回データ工学ワークショップ (DEWS2004) 論文集 6-B-04, 2004.

[4] Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, Vol. 07, No. 1, pp. 76–80, 2003.

[5] James Rucker and Marcos J. Polanco. Siteseer: Personalized navigation for the web. *Commun. ACM*, Vol. 40, No. 3, pp. 73–75, 1997.

[6] Jason J. Jung, Jeong-Seob Yoon, and GeunSik Jo. Collaborative information filtering by using categorized bookmarks on the web. In *INAP*, pp. 343–357, 2001.

[7] G.Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer.* Addison-Wesley Publishing Company, 1988.

[8] Mecab: Yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.jp/.

[9] K.Maruyama, K.Takasuka, Y.Yagihara, M.Satoshi, Y.Shirai, M.Terada. Real-Time Discovery of Currently and Heavily Viewed Web Pages. In *proc. of WEBIST 2006*, pp. 352–359, 2006.