F-032

ユーザの閲覧傾向を用いた有用な未知情報の推薦
Recommendation System for Novelty and Interest based on User Web Browsing History

近藤 司 ‡ Tsukasa Kondo
原田 史子 † Fumiko Harada
島川 博光 † Hiromitsu Shimakawa

1. はじめに

WWW 上に存在する情報量は膨大で，ユーザが自身にとって必要な情報を適切に取捨選択するのは困難であるため，ユーザの情報探索を支援する推薦システムの技術が研究されてきた．推薦システムはユーザの選択行動から興味を推測し，興味に合致する情報を提示する．しかし，ユーザの興味を正確に反映した情報がユーザにとって有用とは限らない．推薦される情報がユーザにとって既知の場合，推薦結果からユーザが新しく得る情報量は少なく，ユーザは推薦結果に満足できない．

2. 有用な未知情報の推薦の必要性

2.1 推薦におけるユーザの満足度向上の試み

推薦システムにおける満足度を向上するために，未知の情報や意外な情報を推薦する研究がされている [1, 2]．文献 [1, 2] のようにユーザの満足度を向上することを目的とした研究では，ユーザにとって未知で有用な情報を推薦することをコンセプトにしている．しかし，これらの研究ではすべてのユーザに対して 1 つの手法で未知で有用な情報を推薦している．どんな未知の情報に有用性を感じるかは，ユーザの主観に左右される．そのため，さまざまなユーザに対して推薦の満足度を向上させるには，ユーザごとにどんな未知の情報に有用性を感じるかを把握しなければいけない．

2.2 閲覧傾向と有用な未知情報

ユーザが有用だと感じる未知の情報には以下の 3 種類が考えられる．

i ユーザが特に好んでいるジャンルの未知の情報

ii ユーザが興味を持つ複数のジャンルの組み合わせが未知性を発生している情報

iii 興味のあるジャンルであればどんな情報でもいい

i を有用だと感じるユーザは特定の情報を追求したいユーザであると考えられる．例えば，興味のある“ サッカー ”に関する情報の中でも“ 日本代表 ”に関する情報に高い興味を示すユーザは，“ サッカー日本代表 ”に関する情報を多く閲覧する．そのため，“ サッカーの日本代表 ”に関する未知の情報に有用性を感じると考えられる．ii を有用だと感じるユーザは自身にとって興味のある異なるジャンルの組み合わせがのゆえに未知性がある情報に有用性を感じるユーザである．例えば，“ サッカー ”と“ 野球 ”の両方に興味があれば，“ サッカーの本田選手と野球選手のイチローが対談 ”のように，2 つのジャンルの組み合わせが発生している情報に有用性を感じると考えられる．iii を有用だと感じるユーザは多様な情報に有用性を感じるユーザである．つまり，興味がある“ サッカー ”に関する未知の情報であれば何でも，有用性を感じるユーザである．ユーザは自身にとって有用だと判断した情報を閲覧していくため，i,ii,iii のいずれかに有用性を感じるユーザは，i,ii,iii のいずれかに関する情報を多く閲覧する傾向があると考えられる．そこで，ユーザの閲覧履歴を解析することで i,ii,iii のどの閲覧傾向を持っているのかを判定できると考えられる．

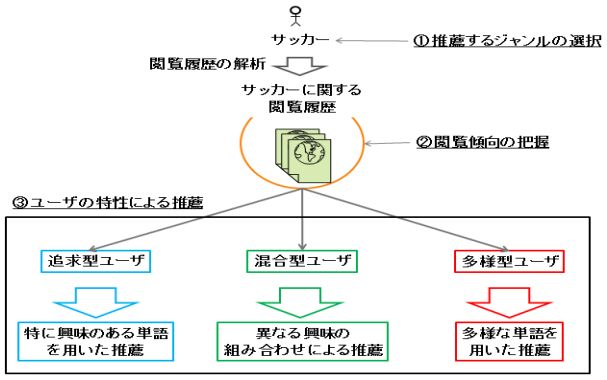

図 1: 提案手法の全体図

3. 閲覧傾向を用いた有用な未知情報の推薦

3.1 ユーザの閲覧傾向に基づく有用な未知情報の推薦

本論文では Web ページを対象として，ユーザの興味のあるジャンルに関する情報を推薦する．ユーザの興味あるジャンルにおいて，どんな未知の情報であれば有用性を感じるのかを判定し，さまざまなユーザに対して有用な未知の情報を推薦することを目的とする．図 1 は本手法の全体図である．本手法には大きく 2 つの処理がある．1 つ目は，ユーザの閲覧傾向を判定する処理である．閲覧履歴を解析し，3 章で述べた閲覧傾向の中でどれを持っているのかを判定する．2 つ目は，推薦処理である．ユーザの特性によって，推薦手法を変更することで，さまざまなユーザに対して有用な未知情報を推薦する．

3.2 推薦ジャンルの選択

本手法ではまず，ユーザに単語を 1 つ選択してもらう．これは，選択された単語に関する Web ページを推薦するためである．次に，選択された単語に関する Web ページ群のみを閲覧履歴から抽出する．ここで，選択された単語に関する Web ページ群を関連 Web ページ群と定義する．閲覧履歴中の各 Web ページ群が関連 web ページかどうかは TF/IDF を用いて判定する．ある関連 Web ページの総単語数を N，選択された単語を c，ある Web ページでの単語 c の出現回数を n，関連 web ページ群の数を R，単語 c が出現する関連 web ページ群の数を r，単語 c の TF/IDF 値を $F(c)$ とおき，$F(c)$ を以下のように定義する．

$$F(c) = \frac{n}{N} \cdot log_2 \frac{R}{r} > \theta_1 \quad (1)$$

選択された単語 c の $F(c)$ が閾値 θ_1 を越える Web ページを，関連 Web ページとして抽出する．

† 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

表 1: 閲覧傾向ごとの特徴

	特定情報への固執	他の興味との関連	興味の多様性
追求型	有り	無し	低い
混合型	無し	有り	低い
多様型	無し	無し	高い

3.3 ユーザの閲覧傾向の判定

関連 Web ページ群から，ユーザの閲覧傾向を判定する．表 1 は 3 章で述べた閲覧傾向の特徴をまとめたものである．表 1 より追求型ユーザは特定の情報を多く閲覧する傾向にあると考えられる．ここで，特定の情報を追求しているかどうかを表す指標を追求度と定義する．特定の情報を多く閲覧すれば，特定の単語が多く出現すると考えられるため，追求度は関連 web ページ群における単語の出現率とする．関連 web ページ群に出現する単語群を $\{w_1,w_2\cdots w_i\}$,$\{w_1,w_2\cdots w_i\}$ の数を i，単語 w_i の追求度を $I(w_i)$ とおくと，追求度は

$$I(w_i) = \frac{w_i}{i} > \theta_2 \tag{2}$$

と表現できる．$I(w_i)$ が閾値 θ_2 を超える単語をユーザが特に閲覧している単語として抽出し，$I(w_i)$ が閾値を超える単語を抽出できるユーザを追求型ユーザとする．

表 1 より混合型ユーザは，ユーザが興味のある異なるジャンルの組み合わせが発生している情報を多く閲覧すると考えられる．ここで，関連 web ページ群においてユーザが興味のある異なるジャンルの組み合わせが発生している割合を混合度と定義する．混合度を計算するためにまず，閲覧履歴から関連 web ページ群を除外した Web ページ群から単語を抽出し，その出現回数を計算する．ここで，出現回数が一定以上の単語群を $\{x_1,x_2\cdots x_j\}$,$\{x_1,x_2\cdots x_j\}$ の数を j とする．$\{x_1,x_2\cdots x_j\}$ は，関連 web ページ群とは異なるジャンルのユーザの興味を表していると言える．関連 web ページ群とは異なるジャンルにも関わらず，関連 web ページ群に $\{x_1,x_2\cdots x_j\}$ が出現した場合，ユーザが興味ある異なるジャンルの組み合わせが発生したと言える．そこで，$\{x_1,x_2\cdots x_j\}$ 中で関連 web ページ群にも共通して出現している単語の割合を計算する．ここで，混合度を O，$\{x_1,x_2\cdots x_j\}$ 中で関連 web ページ群にも共通して出現している $\{x_1,x_2\cdots x_j\}$ の数を k とすると混合度 O は以下で表現できる．

$$O = \frac{k}{j} > \theta_3 \tag{3}$$

O が閾値 θ_3 を越えるユーザを混合型ユーザとする．

表 1 より多様型ユーザは一貫性なくさまざまな情報を閲覧すると考えられる．ユーザの閲覧傾向に一貫性があるかを判定するために，関連 Web ページ群に出現している単語同士の類似度の平均を計算する．関連 Web ページ群に出現している単語同士の類似度の平均が高ければ，個々の関連 Web ページ同士は似通った情報を含んでいると言えるが低ければ，個々の関連 Web ページ同士は似通った情報ではないと言える．関連 web ページ群に出現する単語同士の類似度を S として，関連 Web ページ群に出現しているすべての単語同士の類似度を計算する．

$$S(w_l, w_m) = \frac{w_l \cap w_m}{w_l \cup w_m} \tag{4}$$

$w_l \cap w_m$ は w_l,w_m が同時に出現する関連 web ページ群の数，$w_l \cup w_m$ は w_l,w_m のどちらかが出現する関連 Web ページ群の数である．式 4 は，w_l,w_m が同時に出現している Web ページが多ければ，w_l,w_m の類似度は高いという考えに基づいている．単語同士の類似度を用いて，関連 Web ページに出現しているすべての単語の組み合わせの類似度の平均を求める．単語同士の類似度の平均を AS とおくと

$$AS = \frac{2(\sum_{l=1}^{i}\sum_{m=1}^{i} S(w_l, w_m))}{{}_iC_2} > \theta_4 \tag{5}$$

ただし，$l \neq m$ である．ここで，AS が閾値 θ_4 より低いユーザを多様型ユーザとする．

3.4 閲覧傾向ごとの推薦手法

追求型ユーザは，ユーザが特に閲覧している情報に関する未知の情報を推薦すれば良い．そこで，追求度が閾値 θ_2 を越える単語と共起している単語の中でユーザがあまり閲覧したことの無い単語を用いて推薦をする．関連 Web ページ群において追求度が閾値 θ_2 を越える単語の数を a，その単語群を $\{y_1,y_2\cdots y_a\}$ とする．$\{y_1,y_2\cdots y_a\}$ と関連 Web ページ群において共起している単語を抽出する．例えば，単語 y_1 と関連 Web ページ群において共起している単語群の数を b，その単語群を $\{z_1,z_2\cdots z_b\}$ とする．次に，$\{z_1,z_2\cdots z_b\}$ の関連 Web ページ群と WWW における出現率を計算する．関連 Web ページ群と WWW における出現確率の差が大きい単語と単語 y_1 の組み合わせはユーザにとって未知である可能性が高い．そこで，$\{z_1,z_2\cdots z_b\}$ の関連 Web ページ群と WWW における出現率に対して T 検定をする．$\{z_1,z_2\cdots z_b\}$ の中で関連 Web ページ群と WWW における出現率の差が優位である単語と単語 y_1 の組み合わせを用いて推薦をする．この作業を $\{y_1,y_2\cdots y_a\}$ のすべての単語で繰り返す．

混合型ユーザには，ユーザがまだ閲覧していないユーザの興味あるジャンルの組み合わせで推薦する．そこで，$\{x_1,x_2\cdots x_j\}$ の中で関連 Web ページ群に出現していない単語とユーザが選択した単語 c を用いて推薦する．

多様型ユーザには，ユーザにとって未知である可能性の高い情報を推薦すれば良い．そこで，追求型ユーザと同様に T 検定を用いて未知である可能性の高い単語を抽出する．$\{w_1,w_2\cdots w_i\}$ の関連 Web ページ群と WWW における出現率に対して T 検定をする．関連 Web ページ群と WWW における出現率に優位な差がある単語と 4.2 節でユーザが選択した単語 c を用いて推薦をする．

4. おわりに

本論文ではユーザの興味あるジャンルにおいて，さまざまなユーザに対して有用な未知情報を推薦する手法を提案した．今後は本手法の有用性を評価する実験をする予定である．

参考文献

[1] 村上知子，森紘一郎，折原良平: 推薦の意外性向上のための手法とその評価．人工知能学会論文誌，vol.24, no.5, pp.428-436, 2009 年.

[2] 住元宗一郎，中川博之，田原康之，大須賀昭彦: コンテンツ投稿型 SNS における未知性と意外性を考慮した推薦エージェントの提案．電子情報通信学会論文誌，vol.j94-D, No.11, pp.1800-1811, 2011 年.