6ZK-08

出現単語の組み合わせに基づく記事文に対する興味の強さ推定

山下 慶悟† 遠藤 慶一‡ 黒田 久泰‡ 小林 真也‡
†愛媛大学工学部情報工学科 ‡愛媛大学大学院理工学研究科

1. はじめに

近年インターネットの普及が進み，多くの人がインターネットから情報を手軽に入手することが可能になった．その一方でインターネット上には膨大な量の情報が溢れることになり，その中から自分にとって有用な情報だけを選別する作業はユーザにとって大きな負担となる．この問題を情報過多という．また，年齢などにより，情報サービスの恩恵を受けることができる人とそうでない人の間に格差が生じている．この問題をディジタルディバイドという．この２つの問題を解決することを目的として，個人向け情報配信システム PINOT (Personalized INformation On Television screen) が開発された [1]．PINOT は情報配信サーバから送られてきた記事に対し，ユーザの興味に合わせて選別を自動的に行い，画面に表示し，表示された記事に対するユーザの操作から，記事への関心を類推し，記事を構成する単語に対する興味の度合いを学習するシステムである．

2. 個人向け情報配信システム PINOT

PINOT の構成を図１に示す．PINOT は情報通信サーバ，セットトップボックス，リモコン，テレビで構成されている．

PINOT の課題として，文脈によって意味の変わる単語が記事情報に含まれている場合，ユーザの興味を反映できない可能性があるということが挙げられる．

文脈によって意味の変わる単語の１つの用法に対して高い興味を示した場合，別の用法のときでも興味の度合いの値はそのまま用いられるため，他の単語よりも興味の度合いが高くなることが考えられる．このとき，ユーザにとって興味のない記事情報が「興味あり」と判定される可能性がある．反対に文脈によって意味の変わる単語の興味の度合いが他の単語よりも低い場合，ユーザにとって興味のある記事情報が「興味なし」と判定される可能性がある．

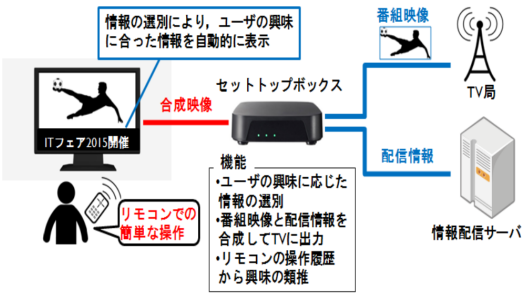

図 1: PINOT の構成

本研究の目的は，記事情報に文脈によって意味の変わる単語がある場合，ユーザの興味を正確に反映できないという問題の解消である．また，本研究の目標は記事見出し文を構成する単語の組み合わせによる興味の学習を行う手法を用いて，ユーザの興味を反映した記事の選別を行い，従来法と比較して提案手法の情報フィルタリング性能を評価することである．

3. 提案手法

本研究では，記事情報に含まれる単語に加え，単語同士の組み合わせについても興味の学習を行い，単語，単語同士の組み合わせの興味の度合いの値の平均値を用いて記事情報に対する関心の度合いを計算するという手法を提案する．

この手法を用いることで，記事情報に文脈によって意味の変わる単語が含まれている場合，ユーザの意思に反した判定が行われる可能性を下げることが期待できる．

3.1. 提案手法における記事情報に対する興味の有無の判定方法

以下の手順で記事情報に対する関心の度合いを計算して興味の有無の判定を行う．

1. 配信された記事情報から単語と組み合わせを抽出
   記事情報から名詞を抽出し，名詞同士で組み合わせを作る．組み合わせのパターンは ${}_MC_2$ 通り (M: 抽出した名詞の数) となる．
2. ユーザプロファイルから単語と組み合わせの興味の度合いを取得
   抽出した単語 $\omega_m (m = 1, 2, 3, \cdots, M)$ と組み合わせ $\omega_{m,n} (n = m+1, m+2, \cdots, M)$ がユーザプロファイルに記録されている場合，ユーザプロファイルを参照して興味の度合い $i(\omega_m)$, $i(\omega_{m,n}) (0 \leq i(\omega_m), i(\omega_{m,n}) \leq 1)$ を取得する．新出単語は興味の度合いの値を 1，新出組み合わせは興味の度合いの値を記事情報に対する興味の有無の判定に用いる閾値と同じ値にする．
3. 記事情報に対する関心の度合いの計算
   記事情報に対する関心の度合い I を計算する．値が大きいほどユーザの記事情報に対する関心が高く，小さいほど興味が低い．

$$I = I_{word} \times \beta + I_{combi} \times (1 - \beta)$$

$$I_{word} = \frac{\sum_{m=1}^{M} i(\omega_m)}{M}$$

$$I_{combi} = \frac{\sum_{m=1}^{M} \sum_{n=m+1}^{M} i(\omega_{m,n})}{{}_MC_2}$$

β は単語の興味の度合いを重視する割合 $(0 \leq \beta \leq 1)$ である．

Estimation of strength of interest in a news sentence by combination of appearing words for the information distribution system

†K. Yamashita
Department of Computer Science, Faculty of Engineering, Ehime University
‡K. Endo, H. Kuroda, S. Kobayashi
Graduate School of Science and Engineering, Ehime University

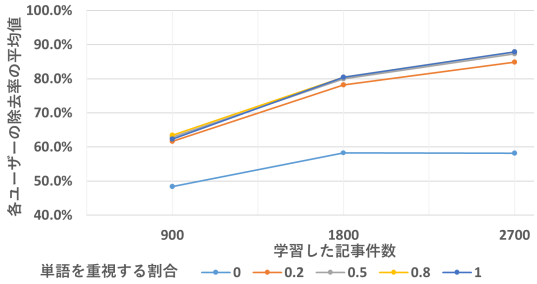

図 2: 学習した記事件数に対する除去率の推移

4. 記事情報に対する興味の有無の判定
   I の値が閾値以上であれば「興味あり」，そうでなければ「興味なし」とする．
5. ユーザプロファイルの更新
   興味の有無の類推結果を基にユーザプロファイルの内容を更新する．

$$i(\omega_m) := \alpha \times i(\omega_m) + (1-\alpha) \times J$$
$$i(\omega_{m,\,n}) := \alpha \times i(\omega_{m,\,n}) + (1-\alpha) \times J$$

α は過去の興味の度合いを重視する割合 ($0 \leq \alpha \leq 1$)，J は興味の有無 (0 または 1) である．

4. 評価

4.1. 実験の手順

提案手法の情報フィルタリング性能を評価するために実験を行う．実験の手順は以下の通りである．

1. 学習に用いる $N(N = 900, 1800, 2700)$ 件の記事見出し文，選別に用いる 200 件の記事見出し文に対して興味の有無の回答を行う
2. 1 で行った記事見出し文に対する興味の有無の回答の内容を入力してユーザごとにユーザプロファイルを作成する
3. 選別に用いる 200 件の記事見出し文をユーザプロファイルを用いて「興味あり」と「興味なし」に選別する
4. 選別した記事見出し文と，それに対する実際のユーザの回答に対して評価基準を用いて評価を行う

3，4 については単語の興味の度合いを重視する割合 $\beta(\beta = 0, 0.2, 0.5, 0.8, 1)$ を設定して行う．

4.2. 評価基準

評価基準には除去率，再現率を用いる．除去率とは，ユーザにとって興味のない記事の中で，記事の選別の際に「興味なし」と判定された記事の割合である．除去率が高いほど，ユーザの興味のない記事を除去できたことを意味する．再現率とは，ユーザにとって興味のある記事の中で，記事の選別の際に「興味あり」と判定された記事の割合である．再現率が高いほど，ユーザの興味のある記事を選別できたことを意味する．

5. 結果と考察

7 名のユーザに対して実験を行った．図 2 に学習した記事件数と各ユーザの除去率の平均値の推移，図 3 に学習した記事件数と各ユーザの再現率の平均値の推移を示す．

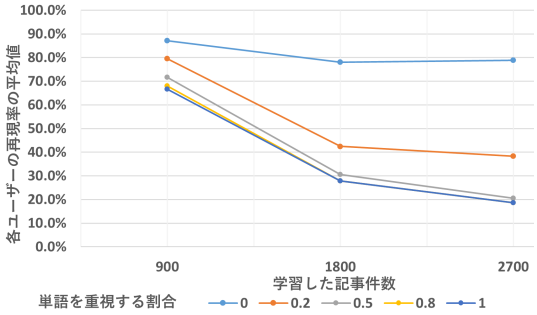

図 3: 学習した記事件数に対する再現率の推移

図 2 より，除去率は組み合わせの興味の度合いのみを用いて選別したとき ($\beta = 0$) に一番低く，それ以外はほぼ同じ値となっている．除去率が $\beta = 0$ で一番低い値になった理由として，新出組み合わせの興味の度合いの値，過去の興味を重視する割合，記事情報に対する興味の有無の判定に用いる閾値が同じ値になっていたことが考えられる．本研究の場合，単語および組み合わせは少なくとも 2 回「興味なし」と学習されなければ，その単語および組み合わせの興味の度合いは閾値を下回らない．また，組み合わせは単語に比べて大幅に数が多く，登場する頻度も低いため，記事の選別を行う際，新出組み合わせや 1 回しか登場していない組み合わせの存在によって記事に対する興味の度合いが閾値未満になりにくい．そのため，選別によって「興味あり」と判定されたが，実際にユーザは興味がない記事が多数存在していたことで除去率が低くなったと考えられる．

図 3 より，再現率は組み合わせの興味の度合いのみを用いて選別したとき ($\beta = 0$) に一番高い値になっていることがわかる．このことから，組み合わせの興味の度合いの値のみを用いて記事に対する興味の類推を行った場合，従来法に比べ，ユーザの興味のある記事を選別できると言える．つまり，ユーザにとって興味のある記事が，文脈によって意味の変わる単語によって「興味なし」と判定される可能性を下げることができると言える．

6. おわりに

本研究では PINOT において記事情報に出現する単語に加え，単語の組み合わせについても興味の学習を行い，そのデータを用いて記事情報に対する関心の度合いを計算するという手法を提案し，実験を行った．その結果から，提案手法を用いることで，ユーザにとって興味のある記事が，文脈によって意味の変わる単語によって「興味なし」と判定される可能性を下げることができるとわかった．今後は，除去率を向上させつつ，再現率を低下させないように，新出組み合わせの興味の度合いの値，過去の興味を重視する割合，記事情報に対する興味の有無の判定に用いる閾値を変化させて研究を行っていく必要がある．

参考文献

[1] 森 健，柏木 紘一，樋上 喜信，小林 真也，“Ticker に対する表示操作履歴に基づいた興味の有無の推論”，情報処理学会，グループウェアとネットワークサービスワークショップ 2005 論文集，111-116