個人属性を考慮した情報フィルタリング

5Z－1

北川 結香子**，中嶋 卓雄*，河北 隆生***，中村 良三*

*熊本大学 工学部，**熊本県立大学 総合管理学部，***熊本県工業技術センター

1　はじめに

コンピュータネットワークの発展により情報洪水と呼ばれる問題が発生している．一方，新聞記事もWeb上で電子化されネットワークで公開されつつある．このような状況の中で，地域の新聞である熊本日日新聞社の協力を得て，新聞記事に対するフィルタリングシステムの開発を行ってきた[1].

このシステムでは処理をフィルタリングおよびフィードバックの2つに分割し，フィルタリングでは，ベクトル空間法に基づき類似度を計算し，閾値を越えたものを抽出し，フィードバックの部分では，関連フィードバック法に基づき処理を実現している．

しかし，キーワードのみでユーザデータと記事データのマッチングをするだけでは，適合率，再現率の向上に限界が生じている．

本稿では，ユーザの個人属性や，個人の特性を考慮に入れてフィルタリングの精度を向上させるフィルタリング処理を提案する．

2　データの分類

ユーザのプロファイル情報を，(1) 明示的な情報，(2) 非明示的な情報に分類し，(1) については、ユーザにキーワードとして入力させ，(2) については，ユーザの個人的な情報とユーザの動作履歴データからその特性を抽出する．

Information Filtering based on The User Characters
Yukako Kitagawa**,Takuo Nakashima*,
Takao Kawakita***Ryozo Nakamura*

*Faculty of Engineering, Kumamoto University,

**Prefectual University of Kumamoto

***Kumamoto Industrial Research Institute

具体的には，ユーザ個人の情報とユーザが所属するグループの情報に分類する．ユーザが所属するグループとは，当面，ユーザの個人的属性である年齢，性別，職業，出身など，属性が独立となるものをグループとして考えている．事前の実験で，これらの中で性別が最も好みに差が現われることが判明しているので，今回は，性別について扱い，性別毎にサンプリングしたデータにより特徴を抽出する．個人の情報は，ユーザの動作履歴データに基づき，キーワードに対する興味および新聞カテゴリーに対する興味の情報を抽出する．ここで，新聞カテゴリーとは新聞の面名に相当するもので，トップ，経済，スポーツ，総合などに分類されているものを利用している．

したがって，ユーザプロファイル情報は以下のキーワード情報，カテゴリー情報から構成する．

2.1　キーワード情報

ユーザのキーワードに対する興味は「関心度」と呼ぶ指標によって管理する．関心度は，0から1までの数値で表す．対象とするキーワード集合として，(1) ユーザが入力したキーワード，(2) ユーザが読んだ記事より抽出したキーワード，(3) 同じ性別のユーザが読んだ記事より抽出したキーワード，から構成する．

2.2　カテゴリー情報

ユーザのカテゴリー集合に対する興味は「興味分布」と呼ぶ指標によって管理する．記事が属するカテゴリーにユーザがどの程度関心を持つのかを表わす指標で，0から1までの数値で表わし，カテゴリー分類の単位ごとに，その数値の合計を1とする．今回は，(1) 新聞カテゴリーへの興味分布，から構成す

る．将来的には，新聞カテゴリーの階層化，ユーザに固有な新聞のクラスタリングによるユーザ固有なカテゴリー興味分布などについて考察するつもりである．

2.3 新聞記事情報

記事から基本辞書を用いて自然言語解析を行いキーワードおよびその頻度を抽出し記事データベースを作成する．

3 フィルタリング

フィルタリングにはベクトル空間法を，フィードバックには関連フィードバック法を適用する．以下ではフィルタリング処理についてのみ詳細化する．

3.1 フィルタリング処理

k個のキーワードが利用できると仮定する．それぞれのキーワード K_iにベクトル V_iを対応させ，k次元のベクトル空間を定義する．このベクトル空間において，記事 A_rを，次のように表現する．

$$A_r = \sum_{i=1}^{k} a_i^r V_i \tag{1}$$

ここで，係数 a_i^rは記事 A_rにおけるキーワード K_iに対する値であり，

$$a_i^r = w * b_i^r + c_i^r \tag{2}$$

b_i^r： キーワード K_iの見出しにおける出現頻度
c_i^r： キーワード K_iの本文における出現頻度
w： 見出し重み

とする．

プロファイル情報 P_qに対しても，次のようにベクトル表現する．

$$P_q = \sum_{i=1}^{k} p_i^q V_i \tag{3}$$

ここで，係数 p_i^qはプロファイル情報 P_qにおけるキーワード K_iに対するユーザの関心度であり，

$$p_i^q = w_i * I_i^q + w_r * R_i^q + w_s * S_i^q \tag{4}$$

I_i^q： ユーザが初期入力したキーワード K_iに対する関心度
w_i： 初期入力キーワードに対する重み
R_i^q： ユーザが読んだ記事から抽出したキーワード K_iに対する関心度
w_r： 読んだ記事から抽出したキーワードに対する重み
S_i^q： 同じ性別のユーザが読んだ記事から抽出したキーワード K_iに対する関心度
w_s： 性別に対する重み

とする．

記事 A_rとプロファイル情報 P_qとの類似度を以下のように定義する．

$$sim(A_r, P_q) = e_c^r * A_r \bullet P_q = e_c^r * \sum_{i,j=1}^{k} a_i^r p_j^q V_i \bullet V_j \tag{5}$$

ここで，係数 e_c^rは記事 A_rが属する記事カテゴリー Cに対するユーザの興味を表わし，

$$e_c^r = w_d * d_c^r \tag{6}$$

d_c^r： 記事 A_rの属するカテゴリー C における興味分布値
w_d： 興味分布に対する重み

とする．

k個の語ベクトル Vはそれぞれ直交していると仮定して，

$$sim(A_r, P_q) = \sum_{i=1}^{k} a_i^r p_i^q \tag{7}$$

となる．

フィルタリングには，この類似度に対する閾値を設定して，閾値を越えた記事のみ出力する．

4 おわりに

本稿では，ユーザの個人属性を考慮したフィルタリング手法を提案した．今後は，評価実験により，パラメータの妥当な値を求めていきたい．

参考文献

[1] 西孝史，中嶋卓雄，北川結香子，河北隆生，中村良三“ユーザの好みを考慮した新聞記事のランキング”，第 55 回情報処理学会全国大会論文集，5Q-2,Vol.3,pp.228-229,1997.