7S-01

個人情報フィルタを使ったエッジ型情報収集支援における個人クエリ間の違いの定量化

芦川拓実† 風尾勇侑† 佐治寿一† 小林暉† 金道敏樹†

†金沢工業大学工学部情報工学科

1 はじめに

エッジ型の情報収集支援技術へ、ユーザの視野を広げる機能を付加することを目指し、端末に個人の情報収集クエリを持つ個人情報フィルタを基礎に通信を介して個人のクエリを柔軟に相互利用できるエッジ型サービスを、我々は構想している。その際、ユーザの視野拡大に効果的な他人の個人クエリとはどのようなものかを知る必要がある。今回、ユーザ間の興味の類似性に注目して、個人クエリ間の違いの定量化方法を検討したので報告する。

2 ユーザプロファイリング技術と興味の推定

エッジ型の情報収集支援においては、端末での計算量、メモリ使用量は少ないことが望ましいので、図1のようなインターフェースを設計し、提示した記事に対するユーザから興味の有無の2値フィードバックを求める個人適応型情報フィルタ INSOP をユーザプロファイリング技術として採用した [1]。

Judge	Movie	Title	Abstract
○ : ×		「○○の冒険」	○○が□□して冒険に出るのだが…
○ : ×		「恐怖の××屋敷」	その屋敷には× ×が住んでいるとされ…
○ : ×		△△が愛に変わるまで	○○はその日、□□に屋上である事を伝えた。…

図 1: 試作情報システムのインタフェースの概念図

個人適応型情報フィルタ INSOP は、提示したコンテンツに対するユーザから興味の有無の2値フィードバックを受けてユーザ個人の興味を獲得し、それに基づいて新しいコンテンツに対するユーザの興味の度合いの推定値の大きいものから推薦/提示する技術である。その個人クエリ（個人プロファイル）は、

user name	P^i	P^u
k_1	Q^i_1	Q^u_1
k_2	Q^i_2	Q^u_2
⋮	⋮	⋮
k_N	Q^i_N	Q^u_N

である。ここで、user name はユーザ識別子、P^i はユーザが興味があると答えた回数、P^u はユーザが興味がないと答えた回数、k_n はキーワード、Q^i_n はユーザがキーワード k_n が付いているコンテンツに対して興味があると答えた回数、Q^u_n はユーザがキーワード k_n が付いているコンテンツに対して興味がないと答えた回数、である。

そして、ユーザ α のキーワード k_n の興味の度合いの推定値は

$$SKC(\alpha, k_n) = q^i(k_n) \log \frac{q^i(k_n)}{p^i} - (1 - q^i(k_n)) \log \frac{1 - q^i(k_n)}{1 - p^i} \quad (1)$$

$$p^i = \frac{P^i}{P^i + P^u} \quad (2)$$

$$q^i(k_n) = \frac{Q^i(k_n)}{Q^i(k_n) + Q^u(k_n)} \quad (3)$$

と見積もられる。

コンテンツ a に対するユーザ α の興味の度合いの推定値 $C(\alpha, a)$ は、そのコンテンツ a に付いているキーワード $k_n(a)$ のユーザ α の興味の度合いの推定値 $SKC(\alpha, k_n(a))$ の総和

$$C(\alpha, a) = \sum_n SKC(\alpha, k_n(a)) \quad (4)$$

で与えられる。

3 視野拡大支援のポイント

前述の各個人情報フィルタの中にある個人クエリ（個人プロファイル）は、対応するユーザの個人視野を表している。したがって、この個人クエリを実効的に広げることができれば、ユーザへより広い視野に基づい

Quantification of differences between personal queries on edge type information retrieval technology supported by a personal information filter
†Takumi Ashikawa,Yusuke Kazao, Hisakadu Saji, Hikaru Kobayashi, Toshiki KINDO
†Department of Information and Computer Science, College of Engineering, Kanazawa Institute of Technology

て情報提供ができると言うのが、我々の構想のポイントである。

そして、個人クエリを実効的に広げる方法の一つは、ユーザ α の個人情報フィルタが興味の似通った他者 β の個人クエリを利用する方法である。

個人クエリの利用のレベルには、

- 個人クエリそのものを利用
- 個人クエリに基づく興味の推定値を利用

の二つのレベルがある。ここでは、後者を採用する。その理由は、ユーザーの検索履歴や行動履歴を大量に収集し分析を行っている Google や Amazon の推薦システムとの技術的な差別化もあるが、個人クエリの中身を見ることなく、かつ少ないサンプル数で個人クエリ間の違いの定量化ができるというメリットがあるからである。

今、ユーザ α,β の個人情報フィルタの出力がコンテンツ a に対する各ユーザの興味の度合いの推定値 $C(\alpha,a),C(\beta,a)$ であるとする。このとき、ユーザ α に対する興味の度合いの推定値を

$$C^{+}(\alpha,a) = C(\alpha,a) + \sum_{\beta} S(\alpha,\beta)C(\beta,a) \qquad (5)$$

と置き換えることにすれば、他者の個人クエリの中身を見ることなく活用できる。

次の問題はここに現れる係数 $S(\alpha,\beta)$、誰の個人クエリを重視し、誰の個人クエリを使わないかを決める係数として何を採用するかである。

4 相関係数を用いたクエリ間の違いの定量化

直観的には、他者の個人クエリの利用がノイズとならないよう似通った興味を持つ他者の個人クエリを重視することが望ましい。この直観に従えば、係数 $S(\alpha,\beta)$ は、似通った興味を持つ他者との間では大きな正の値を取り、正反対の興味を持つ他者との間では負の値を取るような類似性評価値 (違いの評価値) であるとよい。

個人クエリの各キーワード k_a が持つユーザの興味の度合いの推定値 $SKC(\alpha,k_a)$ の値に立ち戻って類似性を評価すれば、高い精度が得られるかもしれないが、それは個人クエリを直接共有するか、センターでの管理を行う必要があり、我々の構想にはそぐわない。

ここでは、精度は少し落ちるかもしれないが、ユーザ α,β の興味の度合いの推定値

$$\{(C(\alpha,a),C(\beta,a)),a=1,2,\cdots\}$$

の相関係数を類似性評価値として採用した。

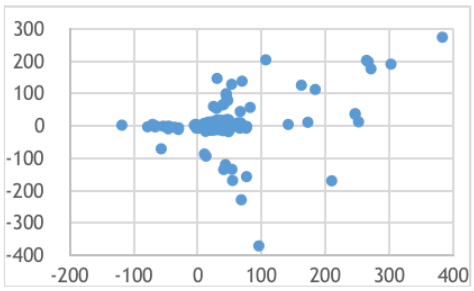

図 2: 異なる人物が同一カテゴリの映画を興味あるものとして選択した場合の散布図 (相関係数=0.53)。

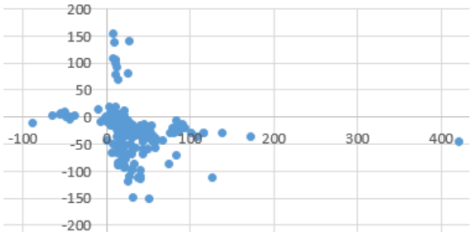

図 3: 異なる人物が SF とラブコメをそれぞれ興味があるものとして選択した場合の散布図 (相関係数=-0.25)。

予備実験として、ユーザに SF、ホラー、ラブコメなどいくつかのジャンルに興味を絞って個人クエリの生成を行い、個人適応型情報フィルタ INSOP の評価値が期待通りの振る舞いをするかを調べたところ、同一のジャンルを志向するユーザ間には正の相関（図 2）が、SF とラブコメのジャンルを志向したユーザ間に最も強い負の相関が見られた (図 3)。また、ホラーとラブコメのように「正反対」のジャンルを志向するユーザ間には負の相関が観測された (相関係数=-0.17)。

これらの結果は、我々の期待（興味の似通ったユーザ間では正の相関、興味の異なったユーザ間では負の相関がある）に合致しており、クエリ間の違いを定量化するために相関係数が有効であることを示唆している。

5 まとめ

個人適応型情報フィルタ INSOP の出力の相関係数によって個人クエリの類似性を評価・利用することで、似通った興味を持つ他者を重視するような個人クエリの相互利用が可能であることを示した。学会においては、より詳細な実験結果の報告を行いたい。

参考文献

[1] Toshiki KINDO *et.al.* Adaptive Personal Information Filtering System that organizes personal profiles automatically,Proceeding of IJCAI97, 1997.