ユーザの好みを考慮した新聞記事のランキング

5Q−2

西 孝史*，中嶋 卓雄*，北川 結香子**，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

近年，コンピュータネットワークが発達するに伴って，その利便さと同時に，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WWW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，ユーザが望む記事を効率よく提供する情報フィルタリングが求められている．

我々は，マルチエージェントによる新聞記事に対するフィルタリングシステムの提案を行ってきた．[1] このシステムでは処理を (1) フィルタリング，(2) フィードバックの2つに分類し，フィルタリング過程において記事のランキングを試みている．

従来，文章のランキングは文書検索の分野において活発に研究されており，単語単位，文字単位の情報に基づき処理されている．

しかし，ユーザが明示していない好みや，ユーザのアクセス履歴などの情報を考慮したランキングは行われていない．

本稿では，新聞記事に対する情報フィルタリングの一部である記事のランキング方式について提案するとともに，これを実験的に評価したので報告する．

2 データモデル

ランキングに利用するデータを，(1) ユーザ情報，(2) 記事情報，に分類し，相互のマッチングによってランキング値を算出する．また，ユーザ情報を，(a) 意識的な情報，(b) 無意識的な情報，に分け，(a) については，ユーザが入力したキーワード情報から算出し，(b) については，ユーザが入力した年齢，性別，職業，出身などの個人情報から算出する．

2.1 キーワード情報

ユーザの好みは，次の関心度を用いて表現する．[関心度] とは，カテゴリとキーワードから構成されるペアにユーザがどの程度興味を持っているかを示す指標．0から1までの数値で表す．ここで，[カテゴリ] は，新聞記事を分類する最も大きな枠組みである．また，[キーワード] は，ユーザが興味を持っている単語とする．

2.2 個人情報

ユーザが入力した年齢，性別，職業，出身などの個人属性情報の集合を個人情報とする．各属性に対して，その値が異なれば異なるユーザ集団に属していると考え，ユーザの興味もその集団に影響を受けていると考える．例えば，職業が学生なら学生集団の考え方に影響を受けていると考える．ここでは，その属性の値毎，影響しているキーワード集合を関連DBとして構成する．

2.3 記事データベース

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を抽出し記事データベースを作成する．

3 ランキング方式

3.1 ランキングの概要

図1にランキングの概要を示す．

Ranking of Newspaper Article based on User Interests
Takafumi Nishi*, Takuo Nakashima*,
Yukako Kitagawa**, Takao Kawakita*** Ryozo Nakamura*

*Faculty of Engineering, Kumamoto University,

**Prefectual University of Kumamoto

***Kumamoto Industrial Research Institute

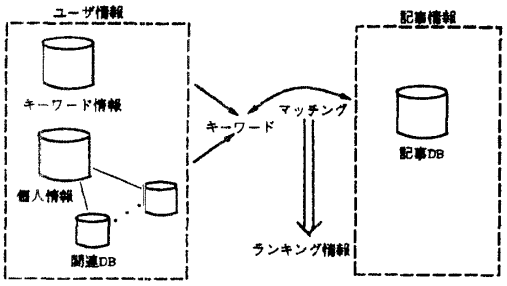

図１：ランキングの概要

3.2　ランキング関数

ランキング値とは，意識的または無意識的にユーザが関心を持つキーワードが記事中に出現する頻度情報に基づく値で，ユーザの記事に対するランキングに相当する．ある記事のランキング値を次のように表わす．

$$\text{ランキング値} = Con * w_c + Uncon * w_u \qquad (1)$$

Con :　意識的な情報に基づくランキング値
$Uncon$:　無意識的な情報に基づくランキング値
w_c :　意識的なランキング値に対する重み
w_u :　無意識的なランキング値に対する重み

ここで，意識的な情報に基づく記事のランキング値を次のように表す．

$$Con = \sum_{i \in \text{keyword set}} (\omega_t * b_i + c_i) * a_i \qquad (2)$$

a_i :　キーワード i に対する関心度 $(0 \leq a_i \leq 1)$
b_i :　キーワード i の見出しにおける出現頻度
c_i :　キーワード i の本文における出現頻度
ω_t :　見出し重み

また，無意識的な情報に基づくランキングでは，属性の値に関連するDBの上位１０個のキーワードに関して，(2)式と同様な関数により求める．なお，関連DB内のキーワードの順位については，システムで初期化すると仮定し，同じグループに属するユーザのアクセスによって更新されるとする．

得られたランキング値を持つ記事を，threshold値によってフィルタリングする．

4　評価

職業が学生である十数名の学生に対し評価実験を行った．[フィルタリングされる記事とユーザが読んだ記事は独立である]と仮定しχ^2検定により検定した．図２に仮定が棄却される有意水準の分布を示す．

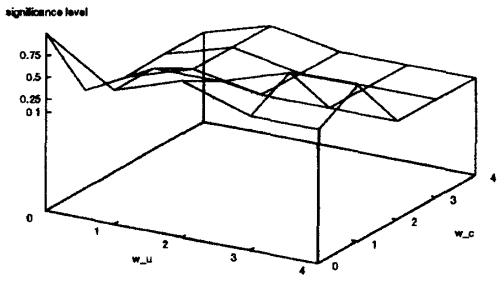

図２：χ^2検定における棄却される有意水準の分布

このデータは，thrshold の値を３とし，見出し重みを３としたときの代表的なユーザに関する実験値である．

この結果から，w_uを２または３に設定したとき，仮定が棄却されやすい，すなわちフィルタリングされた記事と読んだ記事との間に関連があることが示される．

この結果から，threshold の値は３の場合が最も関連があると言えた．また，今回の実験では，関連するキーワードが見出しにほとんど出現しなかったので，見出しの重みに対する考察は不十分であった．

5　おわりに

本稿では，ユーザが関心を持つキーワードを意識，無意識の双方から考慮し，そのデータから記事に対するランキング方式を提案した．また，評価実験により，評価式のパラメータの妥当な値を得ることができた．

今度は，フィードバックを考慮したフィルタリングシステムについて提案をしたい．

参考文献

[1] 中嶋卓雄，稲益康晴，中村良三，河島健一，河北隆生，“新聞記事データベースに対する情報フィルタリング”，情報処理学会，アドバンスド・データベース・シンポジウム'95，pp.113-120，1995