分散作業/学習環境を指向した協調フィルタリングシステムの開発

1S－9

宮原 一弘　　岡本 敏雄

電気通信大学大学院 情報システム学研究科

1. はじめに

ここ数年におけるインターネットの普及にともない，WWW を中心とした情報発信が多くの組織や個人によって行われており，いわゆる情報洪水と呼ばれる状況がインターネットの世界において深刻化しつつある．これに対する解決策の一つとして，情報フィルタリングという技術が注目を浴び，すでに実用的なシステムがいくつか提案されている [1]．また，情報フィルタリングシステムをグループの中で運用し，より効果的な情報収集を支援しようとする協調フィルタリングという概念やシステムも提案されている [2]．

本研究では，大学・企業の研究者や中学校・高等学校の生徒・教員らによって構成される比較的小規模かつ静的なグループにおける協調的な情報収集活動をモデル化し，それに基づいて協調的な情報フィルタリングを実現するシステムの開発を目的とする [3]．

2. 協調フィルタリングシステムの基本設計

本研究では，グループにおける協調的な情報収集活動として，以下のような単純なモデルを考える．

(1) ユーザは WWW ブラウジングなどにより，自主的な情報収集活動を行う．

(2) (1) の過程で，グループ内の他のユーザが興味を持っている（と思われる）情報を得た場合に，該当ユーザに通知する．

本システムでは，このような情報収集活動を，ユーザに煩雑な操作を強いずに実現するために，Shoham のエージェントモデル [4] を採用し，マルチエージェントの協調行動によって，フィルタリングを実現する．システムを構成するエージェントとしては，各ユーザの情報検索・収集過程をモニタすることによりユーザの情報に対する興味領域および情報検索・収集の履歴を管理するユーザエージェント，グループのメンバに共通している興味領域や情報収集履歴を管理するグループエージェントの２種類を構築する．これらのエージェント同士の協調によって，ユーザがどんな情報を欲しているのか，そのためにこれまでどんな情報を検索・利用してきたのか，といったことを相互に把握することが可能となる．図1に本システムのシステム構成を示す．

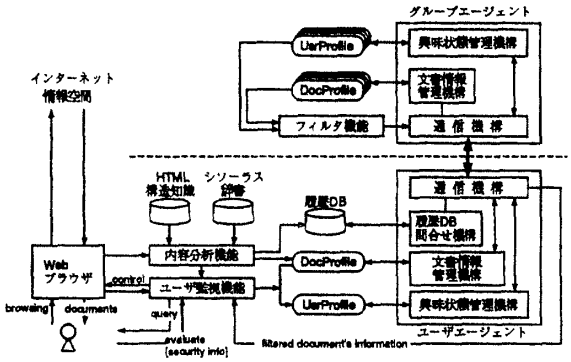

図 1: システム構成

3. フィルタリング機能の実現

3.1 文書プロファイルの生成

本システムでは，文書から抽出したキーワードを分類し，以下の構造を持つ文書プロファイルを構築，フィルタリングに利用する．この過程では，階層構造をなすシソーラス辞書と HTML 文書の構造知識を用いる．

```
DocProfile = {Top word:w,
    {Field words:w},
    {Original words:w} }.
```

Top word は，文書が意味する内容を広範に代表させる語であり，抽出されたキーワードが最も多く出現しているシソーラス階層における最上位のキーワードを採用する．**Field word** は，Top word と同じシソーラス階層に出現するキーワード，すなわち Top word との関連の大きいキーワードの集合によって構成する．**Original word** は，Top word とは直接関連のないキーワード集合により構成する．それぞれのキーワードには，出現頻度および HTML 構造知識から計算される重みが付加される．文書プロファイルの例を以下に示す．

```
DocProfile = { 人工知能:1.3,
    { 知的 CAI:2.4, 定性推論:1.2,..... },
    { グループ学習:1.2, 岡本研:0.6,..... } }.
```

A Development of Collaborative Filtering System Towards Distributed Work/Learning Environment
Kazuhiro MIYAHARA and Toshio OKAMOTO
Graduate School of Information Systems, The University of Electro-Communications
1-5-1, Chofugaoka, Chofu-City, Tokyo, 182, Japan

このプロファイルは，“人工知能”に関する文書であり，それに関連するキーワードとして“知的CAI”，“定性推論”などを含み，人工知能には直接関連のないキーワードとして“グループ学習”，“岡本研”などを含むということを意味している．

3.2 興味プロファイルの表現と獲得

ユーザが情報に持つ興味に関しては，文書プロファイルと同様の構造を持つ興味プロファイルによって表現する．興味プロファイルは複数個持つことが許されるが，Top wordが共通しているものについては，一つのプロファイルに統合して扱う．興味プロファイルの構築はユーザの自主的な情報収集活動をモニタすることによって行われ，そのソースに応じた以下の2通りの手法を用意する．

(1) 検索エンジンを利用した情報収集活動

インターネット上に展開されているWWW検索サービスを利用した個人の情報収集活動にシステムが介入することによって，プロファイルの獲得を行う．処理の手順を以下に示す．

1. 検索エンジンから得られた結果を解析し，すべてのURLに対応するHTMLファイルを収集，キーワードを抽出する．
2. キーワードに関する転置インデックスを生成し提示する．ユーザはリストから自分が必要とする情報を選択し，ブラウジングを開始する．
3. プロファイル生成に関しては，ユーザが検索エンジンへの入力として与えたキーワードをTop wordとする．キーワードを複数与えた場合には，シソーラス辞書を検索し，階層中のより上位に位置する語をTop wordとして採用する．ユーザが転置インデックスから選択した語をField wordとする．Original wordの設定は行わない．

(2) 検索エンジンを利用しない情報収集活動

一方，通常のWWWブラウジング時には，ユーザエージェントがユーザの情報収集活動をモニタし，そこから得られる情報を基にしてユーザと対話を行い，プロファイルを更新していく．処理の概要を以下に示す．

1. ユーザのアクセスしたすべてのHTML文書に対する文書プロファイルを生成する．
2. エージェントはユーザがブラウジングに費している時間を計測し，一定時間以上同一のページを参照していることを認識した場合に，ユーザに対して問い合わせを行う．具体的には，情報が役に立つものであったか，無用なものであったか等をユーザに選択させる．
3. ユーザが文書の有用性を認めた場合にのみ，文書プロファイルの内容に基づいて興味プロファイルの新規作成・更新を行う．

なお，興味プロファイルは，ユーザが明示的に各キーワードを与えて設定することも可能としている．

3.3 プロファイルの比較によるフィルタリング

上記プロセスにより生成された文書プロファイル・興味プロファイルの比較を行うことによってフィルタリング対象となる文書の判定を行う．両者の比較は基本的には，文献検索におけるベクトル空間モデル[5]の類似度に基づいて，以下のように行う．

— Top wordが一致した場合

— Top wordが同一階層内に存在する場合

Field word, Original wordをベクトル空間モデルにおけるキーワードとみなして類似度を計算する．

— Top wordに関連がない場合

フィルタリングの対象外とする．

計算された類似度が閾値より高い情報をユーザの興味に対応する文書とみなし，それに対するポインタ等を通知することによってフィルタリングを実現する．具体的な類似度計算の手法については，現在考案中であり，稿を改めて報告する．

4. おわりに

本稿では，インターネットにおいて，グループによる協調的な情報収集活動を支援する協調フィルタリングシステムについて，システムの概要とフィルタリングの実現手法を中心に述べた．現在はシステムの機能および構成，エージェントの機能・構成を設計中であり，続けてシステムの開発，有効性の評価を行う．

参考文献

[1] 森田昌宏, 速水治夫：情報フィルタリングシステム, 情報処理, Vol.37, No.8, pp.751–758, 1996.

[2] Goldberg, D., et al. : Using Collaborative Filtering to Weave an Information Tapestry, *Comm. ACM*, Vol.35, No.12, pp.61–70, 1992.

[3] 宮原一弘, 岡本敏雄：分散協調作業/学習環境における情報の協調フィルタリング, 信学技法 ET96–94, pp.47–54, 1996.

[4] Shoham, Y. : Agent-oriented Programming, *Artificial Intelligence*, Vol.60, pp.51–92, 1993.

[5] Salton, G. and McGill, M. J. : Modern Information Retrieval, McGraw-Hill, 1983.