知能と複雑系 117−19
(1999. 7. 23)

精度指向のハイブリッド情報フィルタリングの提案

船越要 大黒毅
NTT コミュニケーション科学基礎研究所
〒 619-0237 京都府相楽郡精華町光台 2-4
phone: 0774-93-5240 / email: kf@cslab.kecl.ntt.co.jp

あらまし
情報流通量は増大を続けており, 文献を対象とした情報収集手段として, 高精度の情報フィルタリングが望まれている. 本論文では, 内容に基づくフィルタリングと協調を用いたフィルタリングを統合したモデルを提案し評価を行った. 本モデルでは, 文献は, 内容を示すベクトルと与えられた評価値のベクトルの組で表現される. また, 利用者の特性を行列で扱うことにより, 他の利用者の得意分野と不得意分野を考慮したきめの細かい文献提供を実現する. モデルの性能評価のため, 擬似的に作成した利用者データと文献データを用いてシミュレーションを行った. 評価実験の結果, 既存のフィルタリング方式と比較して精度の向上が確認された.

キーワード　情報フィルタリング, 協調フィルタリング, 利用者プロファイル

A Precision Oriented Hybrid Information Filtering

Kaname Funakoshi, Takeshi Ohguro
NTT Communication Science Laboratories
2-4 Hikaridai, Seika, Kyoto 619-0237
phone: 0774-93-5240 / email: kf@cslab.kecl.ntt.co.jp

Abstract

In this paper, we present a hybrid information (document) filtering model which is a hybrid of the content-based filtering and the collaborative filtering. In the model, each document profile is represented as a pair of a keyword vector and an evaluation vector and each user profile is represented as a matrix of dependency to other users on each keyword. We simulated on a test system with virtual users and virtual documents. The results shows that our model provides appropriate documents to the users with higher precision than other non-hybrid information filtering methods do.

key words　information filtering, collaborative filtering, user profile

1 はじめに

情報流通量は爆発的に増大し続けている．現在，その傾向はインターネットにおいて特に顕著であり[1]，計算機を用いた情報収集 (information search) には，適切な情報を全て提供するだけではなく，利用者の時間の節約のため，無駄な情報の提供を極力減らすことが求められている．本論文では，情報収集手法として情報フィルタリングに着目し，利用者にとって適切な文献を効率的に，特に高い精度で提供することを目的として，文献の内容に基づく情報フィルタリングと，利用者によって文献に与えられた評価を用いた情報フィルタリングを統合したモデルを提案する．

1.1 情報フィルタリング

情報フィルタリング (information filtering) は，流通している文献 (情報流通単位) 集合中から，利用者の広い興味に合致する文献を選択する，あるいは，利用者の興味に合致しない文献を除去 (filter out) することによる情報収集手法である [3].

情報フィルタリングに相当する情報サービスは比較的古くから存在している．カレントアウェアネス (current awareness) や Selective Dissemination of Information と呼ばれるサービスでは，予め登録された利用者の興味や属性に従い，様々な情報源から選択された適切な記事が配布される [8].

1.2 情報フィルタリングの形態

情報フィルタリングの形態としては，内容に基づくフィルタリング (content-based filtering) や協調フィルタリング (collaborative filtering) などが提案されている．

内容に基づくフィルタリングとは，文献の意味内容に着目した手法である．利用者のプロファイル (profile) は，情報検索での質問式 (query) と同じ形式を持ち，利用者の興味を示すキーワードの集合，論理式あるいは重み付けされたベクトルで表現される．この利用者のプロファイルと，個々の文献が持つ，その内容を表すキーワード集合との適合度をマッチング (matching) によって計算し，その結果に基づいて情報提供が行なわれる．

内容に基づくフィルタリングでは，意味内容に沿った文献収集が可能であり，適切に索引づけられた (文献内容を表すキーワードを索引語 (index) として付与された) 文献を，興味を適切にプロファイルに記述した利用者へ提供する，といった目的には優れている．しかし，文献に対する適切な索引づけは専門の索引付与者にとっても困難であるうえ，索引は文献の意味内容を表現しても，その文献がどのような評価を得ているかまでは保証しない．

これに対し，協調フィルタリングは，推薦システム (recommender) とも呼ばれ，文献の内容には関らず，文献に対して他者から与えられた評価に着目した手法である．典型的には，各利用者に対し，似た興味を持つ他の利用者から高い評価を得ている文献の推薦がフィルタリングとみなされる．協調フィルタリングは，コミュニティ内の利用者間の協調活動を利用しており，言い替えると他者の知識を用いた情報収集手法である [5].

協調フィルタリングは，他の利用者の評価に基づくため，既に高い評価を得ている文献の入手には適している．また，索引づけが不要であるため，索引語の付与が特に困難な文献 (画像や音声) の収集には，内容に基づくフィルタリングに対して圧倒的に優る．反面，他の利用者の評価を待たねばならず，文献がフィルタリング可能になるまでの時間がかかる．更に，文献の意味内容は保証されない．

このように，内容に基づくフィルタリングと協調フィルタリングにはそれぞれ得失があり，相互に補完し合えるものと考えられる．

本論文では，内容に基づくフィルタリングと協調フィルタリングの機能を統合したフィルタリング手法を提案する．従来の統合は，内容に基づくフィルタリングと協調フィルタリングを独立に行って結果を合成する方法が取られてきた．本提案手法の特徴は，文献に付与される索引語と利用者による評価とを複合して同時に扱うことである．この手法により，きめの細かいフィルタリングシステムが構成可能となり，精度の高い (無駄な文献提供を極力押さえた) 情報収集が可能となるものと期待される．また，文献に

[1] goo では約 1 億 2 千万の URL がデータベースに含まれている ("http://www.goo.ne.jp/help/engine.html").

索引が付与されていなくとも，また他者による評価を待たずとも，1段階で文献収集が可能となる．

2 従来型情報フィルタリングのモデル

2.1 従来型情報フィルタリングの一般モデル

情報フィルタリングモデルは，情報検索システムと同様，利用者集合，文献集合，マッチング機能，フィードバック機能からなる．以下，利用者のプロファイルの集合を P，文献のプロファイルの集合を D とする．

一般的に従来型の情報フィルタリングのベクトル空間モデルでは，文献と利用者のプロファイルは共に多次元ベクトルで表現され，利用者にとっての文献の適合度は，文献と利用者のプロファイルベクトル間の類似度に基づくマッチングによって求められる [9]．このとき，利用者のプロファイルは，内容に基づくフィルタリングも協調フィルタリングも形式上は全く同一である．

利用者 i のプロファイルをベクトル $\boldsymbol{w}_i$，文献 j のプロファイルを利用者のプロファイルと同次元のベクトル $\boldsymbol{d}_j$ と表現すると，マッチング機能は，ベクトル間の類似度を表すマッチング関数 $match : P \times D \rightarrow \mathbb{R}$ によって実現され，利用者にとっての文献の適合度を表す．例えばcosine measure と呼ばれる手法では，$match(\boldsymbol{w}_i, \boldsymbol{d}_j) = \cos\theta_{(\boldsymbol{w}_i, \boldsymbol{d}_j)}$ という式で表現される．マッチング関数により高いスコアが付けられた文献が，フィルタを通ったものとして，利用者に提供されることとなる．利用者は提供された文献が自らの好みに合ったものかどうかを判定し，その結果をシステムにフィードバックする．フィードバック機能は，1回のフィルタリングが終了するたびに，利用者からフィードバックを受け，利用者や文献のプロファイルや，マッチング機能の更新を行う．

2.2 内容に基づくフィルタリングと協調フィルタリング

内容に基づくフィルタリングでは，利用者及び文献のプロファイルは，各々キーワードに対する重みベクトルとして与えられる．キーワードの総数を N_{term} としたとき，利用者 i および文献 j のプロファイルは，それぞれ N_{term} 次元のベクトル

$$\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iN_{term}})$$

(利用者のキーワードに関する選好),

$$\boldsymbol{d}_j = (t_{j1}, \ldots, t_{jN_{term}})$$

(文献の持つ索引語)

として表現される．また，利用者が提供された文献を評価すると，次回からの文献提供をより適切に行うため，関連フィードバックによって利用者のプロファイルが更新される [6]．

協調フィルタリングでは，文献のプロファイルは，各利用者の評価を表すベクトルとして与えられ，利用者のプロファイルは，他の利用者との類似度で与えられる．システムによって利用者のプロファイルの生成方法は異るが，利用者の興味を示すベクトル同士の類似度に基づくもの [10] や，利用者の文献の推薦の傾向からプロファイルを作成するもの [7] などが存在する[2]．利用者の総数を N_{user} としたとき，利用者 i および文献 j のプロファイルは，それぞれ N_{user} 次元のベクトル

$$\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iN_{user}})$$

(他の利用者の評価に対する依存度),

$$\boldsymbol{d}_j = (t_{j1}, \ldots, t_{jN_{user}})$$

(文献に対する各利用者の評価)

として表現される．ただし，フィードバックは，内容に基づくフィルタリングとは異なり，利用者のプロファイルに対してだけではなく，文献のプロファイルに対しても行われる．つまり，利用者が文献を評価する行動自体が文献のプロファイルに対するフィードバックとなり，その後，システムが利用者のプロファイルを更新する．

概念モデル上は，内容に基づくフィルタリングと協調フィルタリングの本質的な差は，フィードバック機能における文献のプロファイルの扱いにある．情報検索と同様，内容に基づくフィルタリングでのフィードバックは，利用者のプロファイル (情報検索では質問式) のみを更新する．協調フィルタリングでは，文献のプロファイルは利用者の評価であり，フィードバックは利用者のみならず，文献のプロファイルにも作用する．つまり，協調フィルタリングは，文献が動的なプロファイルを持つ情報検索システムと同等である．

[2] ここではモデルの説明の便宜上 $\boldsymbol{w}_i$ をプロファイルと呼ぶが，[7] や [10] では，利用者の特徴を表現するベクトルをプロファイルと呼び，$\boldsymbol{w}_i$ はそれらのベクトルから生成される．

3 本研究のフィルタリングモデル

3.1 行列によるプロファイルを用いたフィルタリング

本論文で提案するフィルタリングにおいては, 高い精度のフィルタリングを実現することを目的とし, 内容に基づくフィルタリングと協調フィルタリングを同時に実現するために, 上記従来型一般モデルを拡張する.

まず, 文献 j のプロファイルは2つのベクトルの組 $\langle \boldsymbol{kw}_j, \boldsymbol{ev}_j \rangle$ で記述する. $\boldsymbol{kw}_j$ は文献の意味内容を表現する固定されたベクトルであり, $\boldsymbol{kw}_j = (kw_{j1}, \ldots, kw_{jN_{term}})$ と表現される. 一方 $\boldsymbol{ev}_j$ は, 各利用者がこの文献に対して与えた評価を表すベクトルであり, $\boldsymbol{ev}_j = (ev_{j1}, \ldots, ev_{jN_{user}})$ と表現される. ただし, $\boldsymbol{kw}_j$, $\boldsymbol{ev}_j$ の各要素は共に実数値を取る.

本研究での手法を提案する前に, この形式の文献とマッチさせるための利用者 i のプロファイルも, 同様に2つのベクトルの組で記述することを考えてみよう. この場合, 利用者の興味をキーワードとして表現したベクトルと, 他の利用者に対する依存度を表現したベクトルの組で記述することになるだろう. このとき, 文献と利用者のキーワードベクトルの類似度は, 内容に基づくフィルタリングとして働き, 文献に与えられた評価のベクトルと利用者の他者への依存度のベクトルの類似度は, 協調フィルタリングとして働き, キーワード, 評価それぞれのベクトルの類似度計算を行い結果の和を求めれば, 確かに文献の内容と他者の評価の双方を用いたフィルタリングとなる.

しかし, 他の利用者には得意とする分野と不得意とする分野がありうる. ベクトルの組では, 分野に関する (内容に基づく) マッチングと利用者に関する (協調による) マッチングは独立に行われ, 各利用者が異る得意分野を持つという関係は表現できない. つまりこの方法では, 各利用者の得意分野に関する特色を生かすことができず, 結果として, 高い精度が期待できないと考えられる.

そこで, 本論文では, 内容に基づくフィルタリングと協調フィルタリングのプロファイルを統合して, 効率的なフィルタリングを達成するための利用者のプロファイルを提案する.

そのために, 利用者 i のプロファイルを2次元に拡張し, $N_{user} \times N_{term}$ の行列 W^i で表す. この形式のプロファイルは, 他の利用者の評価に対する依存度を, キーワードによって変化させることが可能となる表現である. これにより, 内容に基づくフィルタリングと協調フィルタリングとを個々に利用しただけでは実現できない, 利用者の得意分野を考慮したきめの細かいフィルタリングが構成できる.

更にこの形式のプロファイルでは, 評価が全く与えられていない文献に対しては, 純粋に内容に基づくフィルタリングとして動作でき, キーワードが全く与えられていない文献に対しては, 純粋な協調フィルタリングとして動作できるという利点も併せ持つ.

3.2 フィルタリングの動作

具体的には, 利用者 i のプロファイルは以下の形式を持つ.

$$W^i = \begin{pmatrix} w^i_{11} & \cdots & w^i_{1N_{term}} \\ \vdots & & \vdots \\ w^i_{N_{user}1} & \cdots & w^i_{N_{user}N_{term}} \end{pmatrix}$$

ここで, $w^i_{jk} \in \mathbb{R}$ は, 利用者 i にとっての, 他の利用者 j のキーワード k に関する依存度を示す. これは例えば, 利用者 "亀井さん" の, キーワード "Java 技術" に関する評価は信用できる (できない) といったことを意味する.

初期状態の与え方はいくつか考えられる. たとえば, 利用者の初期状態での興味をキーワードとして表現したものが初期化ベクトル $\boldsymbol{v}^i \in \mathbb{R}^{N_{term}}$ として与えられたとすると, 全ての $j(1 \leq j \leq N_{user})$ に対して $w^i_{jk} = v^i_k$ と初期化すれば良い. このとき, 誰も文献を評価していない初期状態では, 初期化ベクトル $\boldsymbol{v}^i$ と文献のキーワードベクトルに関する純粋に内容に基づくフィルタリングと見做せる. 同様に, 初期状態での他の利用者に対する依存度がベクトルとして与えられた場合は, 純粋な協調フィルタリングとして開始することが可能であり, 更には, 全ての要素が同一の値を持つ行列から開始しても構わない. この場合は, 利用者プロファイルは, 興味のある文献の内容に関しても他者への依存度に関しても全くの白紙の状態からスタートすることになる.

マッチング機能は，利用者のプロファイルと文献のプロファイルから実数値を返すマッチング関数によって実現される．この関数は直感的には，上述の例に従うと，「“亀井さん”が高く評価した“Java技術”に関する文献」に高い(低い)スコアをつけるような関数である．マッチング関数は，ここでは次の式を用いる．

$$\begin{aligned} & match(W^i, \langle \boldsymbol{kw}_j, \boldsymbol{ev}_j \rangle) \\ &= \boldsymbol{ev}_j W^i \boldsymbol{kw}_j^t \\ &= \sum_k^{N_{user}} \sum_l^{N_{term}} W_{kl}^i kw_{jl} ev_{jk} \end{aligned} \quad (1)$$

フィードバック機能は，$\boldsymbol{ev}_j$ を書き換えるものと，W^i を書き換えるものの2つが必要になる．ここで，$\boldsymbol{ev}_j$ に対するフィードバックは，文献の提供を受けた利用者が，その文献が適切であるか否かを判断して行う文献のプロファイルに対するフィードバックである．また，W^i に対するフィードバックは，システムが，利用者からのフィードバックを受けてその利用者のプロファイルを書き換える，利用者に対するフィードバックである．

具体的なフィードバック機能は以下のようになる．

まず，文献 j に対する利用者 i による評価が $feedback_{user}(i,j) = e_{ji}$ として与えられる．ここで，$e_{ji} \in \{good, bad\}$ であり，$good$ および bad は，利用者による文献の評価を表す $good > bad$ なる2つの実数である．このとき，文献に対するフィードバックにより，文献のプロファイル中の $\boldsymbol{ev}_j$ ベクトルが，$\boldsymbol{ev}'_j = (ev_{j1}, \ldots, e_{ji}, \ldots, ev_{jN_{user}})$ のように書き換えられる．つまり，文献 j の利用者 i による評価 ev_{ji} が e_{ji} に置き替わる．

次に，利用者のプロファイルに対するフィードバックが行なわれる．利用者に対するフィードバック機能は，一般にフィードバック関数 $feedback_{sys} : P \times D \rightarrow P$ として記述できる．この関数は，現時点の利用者のプロファイルと，利用者の評価が新たに反映された文献のプロファイルから，利用者のプロファイル行列 W^i を書き換える関数である．$feedback_{sys}$ としては種々のものが考えれらるが，ここでは，以下の式を用いる．

$$W^{i\prime} = W^i + \delta_{good} \sum_{ev_{ji}=good} \boldsymbol{x}_j^t \boldsymbol{kw}_j - \delta_{bad} \sum_{ev_{ji}=bad} \boldsymbol{x}_j^t \boldsymbol{kw}_j \quad (2)$$

ここで，δ_{good} および δ_{bad} は定数であり，ベクトル $\boldsymbol{x}_j$ は，文献 j の利用者による評価ベクトル $\boldsymbol{ev}_j$ の要素を，以下に従って整数へ変換したものである．

$$x_{jk} = \begin{cases} 1, & \text{if } ev_{jk} = good \\ 0, & \text{otherwise} \end{cases}$$

つまり，式 (2) は，文献に対して与えられたキーワードに関する，利用者が高く評価した文献に対して高い評価を与えていた評価者に対する依存度を高くし，利用者が低く評価した文献に対して高い評価を与えていた評価者に対する依存度を低くする．

4 実験

モデルの実用性を検証するため，実験システムを構築し，シミュレーションを行った．行列を用いた利用者のプロファイル方式の利点が達成されるかどうかを試験する為の実験である．実験システムは，以下のアルゴリズムで動作する．

アルゴリズム

0. 新しい文献がシステムに入る．
1. その文献をリストに追加．
2. 十分古い文献があればリストから除去．
3. 各利用者について，

3-1. リスト内の全文献に対しマッチング関数の値を計算して保存．

3-2. 計算結果順位づけて利用者に提供．本実験では，上位 20% の文献を提供した[3]．

3-3. 提供した文献に対する利用者からの評価を受けてフィードバックを行い，文献の評価ベクトルと利用者のプロファイルを更新．

4. 0 に戻る．

[3] この値は，現実の利用者にとって役に立つであろうとして恣意的に決定した．全ての文献を提供しその全てに対してフィードバックを受け付ける予備実験を行ったが，フィードバックの効果の速度向上以外には結果の差は認められなかった．

初期状態では, 利用者のプロファイル W^i は, 利用者の興味を示す特徴ベクトル $\boldsymbol{v}^i$ を準備し, 全ての j に関して $w^i_{jk} = v^i_k$ とする. 文献のプロファイルに関しては, 内容を示すベクトル $\boldsymbol{kw}_j$ は予め与えられているものとし, $\boldsymbol{ev}_j$ は, 一様に初期値 ϵ を持つこととする. ここで, ϵ は, 2つの値 $good$ と bad に関して, $\epsilon = \frac{good+bad}{2}$ とする.

マッチング関数は, 式 (1) に従う. 文献 j に対する利用者 i のフィードバックは, 利用者のフィードバック関数 $feedback_{user}$ によってなされる. 利用者のプロファイルに対するフィードバックは, 式 (2) を用いる. δ_{good} と δ_{bad} の値を決定するための予備実験として, δ_{good} および δ_{bad} の値を変更して調査した. 利用者のプロファイルとしてベクトルを持つ方式では, δ_{good} と比較して δ_{bad} が等しいか大きい場合, フィードバックの効果 (すなわち精度の向上) がほとんど観察されず, $\delta_{good} > \delta_{bad}$ の場合にフィードバックの効果が強く見られた. そのため, 本実験では $\delta_{good} = 0.10$, $\delta_{bad} = 0.01$ とした.

ここで, 利用者による評価をシミュレートするため, 正解集合を導入する. 正解集合は文献集合 D の部分集合であり, ある利用者にとって, 本当に興味のある文献の集合である. 「本当に」興味のある文献とは, 文献の内容および質の面において, 利用者の興味を満足させる文献のことであり, 現実の人間を使用しなければ確定することは本来不可能なものである.

全ての利用者に対して, D に対する正解集合を予め作成しておくことにより, 利用者からのフィードバックを模することが可能になる. $feedback_{user}$ の値は, 文献が利用者の正解集合に属しているとき $good$, そうでない場合は bad となる.

4.1 実験データ

実験データとして, 自動的に作成した仮想的な利用者と, 仮想的な文献を準備した. 利用者は実際に文献選択の基準とする「本当のプロファイル」と呼ぶべき行列を持ち, 文献が自分にとって適切であるか否かに対して一貫した評価を行うこととする. 「本当のプロファイル」はランダムに生成するが, キーワードと利用者に対する依存度はある程度重みをつけて生成する. これは, 現実の利用者の興味はある程度偏っているためである. また, 依存度を全てのキーワードと利用者に関して一様に分布させると, 内容に基づくフィルタリング, 協調によるフィルタリングのいづれか一方のみを使用した場合, ユーザの依存度がキーワードや推薦者によらず一様になってしまい, 満足な文献提供が不可能であるためである. 本実験では, 依存度の半数は全てのキーワードと利用者の各々5分の1に集中させることとする.

正解集合は, ランダムに生成した文献と利用者の「本当のプロファイル」のマッチングによって毎回作成する. 新しい文献が入力したときのマッチングは, 全利用者に対して同時に行われ, そこで行われた利用者による文献に対する評価は次回のマッチングに反映される.

利用者数は 100 人, キーワード数は 100 語とした. 実行段階では, プロファイルの初期状態はランダムに決定する. また, 1 回の処理で対象となる文献は, 100 個に固定する.

4.2 評価

フィルタリング性能の評価は, 精度 (precision) と再現率 (recall) の2つの尺度を用いて行う. ここで精度とは, 適合率とも呼ばれ, 利用者に提供された文献集合に含まれている文献の内, 実際に正解集合にも含まれている文献の割合であり, 再現率とは, 正解集合に含まれている文献中の内, 実際に提供された文献の割合である.

情報フィルタリングとしての本来の性能評価として, フィードバックを重ねるに従い, この両尺度が向上していく様子を調べる. 一般に精度と再現率はトレードオフの関係にあるが, ここではより精度の高いフィルタリングが実現できているかどうかに主眼を置く. すなわち, フィードバック回数に対して, 精度が早い段階で高い状態へ達し, そのまま安定する状態が望ましい.

文献を順番に1つづつ入力して処理を行った. 言い換えれば, 100 個の処理対象の文献の内, 1 文献だけが時間の変化に従って入れ替わる.

比較対象として, 内容に基づくフィルタリング, 協調フィルタリングおよび, 内容に基づくマッチングと協調に基づくマッチングの単純な和を用いたフィル

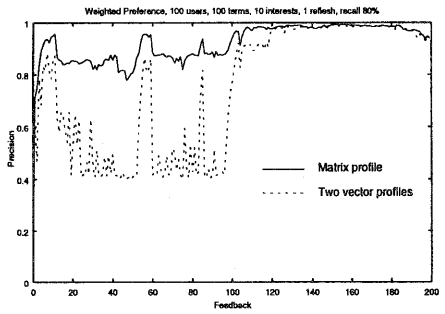

図 1: フィードバック回数に対する再現率 80% での精度の推移

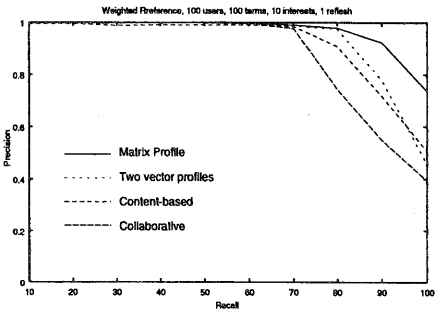

図 2: 120 回のフィードバック以降の再現率あたりの精度の平均

タリング (3.1節で説明したもの) を準備した.

図 1は, 再現率を 80% に固定して観察したときの精度の推移を表す (フィードバックは 200 回まで). ここで, "Matrix profile" とは, 本提案手法での結果を示し, "Two vector profiles" とは, 文献がキーワードのベクトルと他の利用者への依存度のベクトルの組を持つ方式の結果を示す. 本提案手法が, ベクトルの組を持つ方式と比較して, 初期の段階から高い精度を発揮し, そのまま安定していることが分かる.

図 2は, 再現率あたりの精度を表す. ただし, 図 1では両方式共に高い性能を発揮しているように見える 120 回目のフィードバック以降のグラフである. 参考のために, 純粋に内容に基づくフィルタリング (図中では "Content-based" と表示する) および純粋な協調フィルタリング ("Collaborative" と表示する) を用いた場合の結果も示す. 精度と再現率は一般に両立が難しく, 再現率を上昇させると精度は悪化するものであるが, 本提案の手法は, 高い再現率でも, 他の手法と比較して高い精度を実現していることが分かる.

4.3 処理時間

行列を使用したフィルタリングは, ベクトルによるフィルタリングに比較すると極めて大きな計算量を必要とする. 実際にどの程度の処理時間を必要とするかを調査した. 具体的には, 利用者数を 100, キーワード数を 1000, 文献数を 1000 として, 全ての利用者に対する 1 回のフィルタリングの時間を計測した. これは Sun Ultra2 (UltraSPARC 200MHz) 上で 45 分ほど消費する. 各利用者に対するフィルタリングは分散して行なうことが自然であるため, 1 人の利用者に対する消費時間は 30 秒未満であると計算できる. 本研究が想定している利用者集合は, ある程度小規模なコミュニティを対象としており, また情報検索ほど高速な処理は求められないため, この処理時間は実用上妥当なものであると考えられる.

5 関連研究

内容に基づくフィルタリングと協調フィルタリングの統合に関する研究としては, Fab システムが存在する [1]. Fab では, 内容に基づいて文献を収集する collection agent 群と, collection agent 群が収集した文献集合中から各利用者にとって適した文献を選択し提供する selection agent 群の 2 段階によって文献が収集される. 利用者は selection agent に, 文献提供が適切であるかどうかをフィードバックし, selection agent は collection agent に同様にフィードバックする. 全体として, 参加者が共同で collection agent の性能を向上させている協調フィルタリングとして動作する.

その他の内容に基づくフィルタリングと協調フィルタリングの統合に関する研究としては, Basu らによるものがある [2]. Basu らは, 映画を対象とした推薦システムに関して, 映画に予め与えられた属性値か

ら[4], 利用者の嗜好 (好みであるか好みでないか) への関数の学習を扱っている. すなわち, 学習によって文献 (映画情報) を利用者にとって適切に提供する手法が提案されている. 本研究では利用者や文献のプロファイルをフィードバックによって変化させているが, Basu らは, プロファイルへのフィードバックは行わず, マッチング関数自体を変化学習させている.

また, Delgado らは, 内容を考慮した協調フィルタリングを扱っている [4]. Delgado らも機械学習の文脈で研究を進めており, 文献の自動分類や学習方法に着目している.

6 おわりに

本論文では, 比較的小規模のコミュニティを対象とした, 文献内容と利用者からの評価を考慮した情報フィルタリングの方式を提案した.

本研究では, 利用者の, 他の推薦者の各キーワードに関する依存度は常に一貫しているとの仮定に基づいた実験を行ったが, 現実には利用者の興味は必ずしも一貫していない. 本提案の方式が, 利用者の興味が途中で変化する場合にも対応できるものであるかどうかは現在の検討課題である.

また, 現実世界の文献と利用者による実験およびスケーラビリティを考慮した実験も行う予定である.

参考文献

[1] M. Balabanović and Y. Shoham. Content-based, collaborative recommendation. *Communication of the ACM*, vol. 40, no.3, pp.66–72 (1997).

[2] C. Basu, H. Hirsh and W. Cohen. Recommendation as classification: using social and content-based information in recommendation. *Proceedings of AAAI-98*, pp. 714–726 (1998).

[3] N. J. Belkin and W. B. Croft. Information filtering and information retrieval: two sides of the same coin? *Communication of the ACM*, vol.35, no.12, pp.29–38 (1992).

[4] J. Delgado and N. Ishii and T. Ura. Content-based collaborative information filtering: actively learning to classify and recommend documents. M. Klusch and G. Weiß(eds.) *Cooperative information agents II, 2nd International Workshop, CIA'98 Proceedings.* LNAI 1435, Springer-Verlag, 1999, pp.206–215.

[5] D. Goldberg, D. Nichols, B. M. Oki and D. Terry. Using collaborative filtering to weave an information Tapestry. *Communication of the ACM*, vol.35, no.12, pp.61–70 (1992).

[6] D. Harman. Relevance feedback and other query modification techniques. W. B. Frakes and R. Baeza-Yates (eds.) *Information retrieval: data structures & algorithms.* Prentice Hall, 1992, pp.241–263.

[7] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon and J. Riedl. Applying collaborative filtering to Usenet news. *Communication of the ACM*, vol.40, no.3, pp.77–87 (1997).

[8] R. R. Korfhage. *Information strage and retrieval.* Wiley Computer Publishing, 1997.

[9] G. Salton and M. J. McGill. *Introduction to modern information retrieval.* McGraw-Hill, 1983.

[10] U. Shardnand and P. Maes. Social information filtering: algorithms for automating "word of mouth". *Proceedings of CHI'95*, pp.210–217.

[4] Basu らは Internet Movie Database において半構造化されて提供されている映画情報を用いた ("http://www.imdb.com/").

盛光印刷所