

協調フィルタリングに関する研究動向

情報システム学専攻ロボティックス講座

9751090 福原 知宏

tomohi-f@is.aist-nara.ac.jp

1998年 3月 7日

Abstract

ネットワークの普及に伴い WWW や Netnews からの情報獲得の困難性が問題となっている。これに対し新たな情報収集法として協調フィルタリング (Collaborative Filtering) が提案されている。協調フィルタリングは他者の知識を用いる情報収集法である。本稿では協調フィルタリングに関する現在の研究動向と今後の課題について述べる。現在行なわれている研究対象としては、明示的情報に基づくユーザの特徴把握、ユーザ間の類似度計算、個々のユーザに応じた評価予想がある。今後の研究課題としては、暗示的情報を用いたユーザの特徴把握、機械学習、マルチエージェントシステム、プライバシーの保護がある。以下、協調フィルタリングの概念、これまでの研究動向、研究の焦点、今後の課題について述べる。

目次

1	はじめに	5
1.1	ネットワークからの情報収集の困難性	5
1.2	現在の情報収集手法の問題	6
1.2.1	情報検索	6
1.2.2	情報フィルタリング	7
1.2.3	ブラウジング	8
1.2.4	現在の情報収集手法の問題	8
2	協調フィルタリングとは	10
2.1	定義	10
2.1.1	アーキテクチャ	11
2.2	関連概念	13
2.2.1	情報検索と情報フィルタリング	14
2.3	これまでの研究	15
2.3.1	Tapestry	16
2.3.2	GroupLens	16
2.3.3	Ringo	17
2.3.4	Fab	17
2.3.5	PHOAKS	19
3	協調フィルタリングにおける研究対象	20
3.1	ユーザモデリング	20
3.1.1	情報検索におけるユーザモデリング	20
3.1.2	行動履歴に基づくユーザモデリング	23
3.2	類似度計算	25
3.2.1	GroupLens における類似度計算	25
3.3	予想評価の精度	26
3.3.1	Ringo における予想評価の精度	26

4	今後の研究課題	30
4.1	ユーザモデリング	30
4.2	機械学習	31
4.3	マルチエージェントシステム	31
4.4	プライバシーの保護	33
5	おわりに	34
A	WWW 上の情報源	38

目 次

1.1	Information Retrieval	6
1.2	Information Filtering	7
1.3	Browsing	8
2.1	Collaborative Filtering	11
3.1	Oddy の連想グラフ (Ellis の文献より)	21
3.2	Summary of Results	28

表 目 次

2.1	User profile の例	12
2.2	Recommendation の例	12
2.3	情報検索と情報フィルタリング	14
3.1	Feminist の典型	22
3.2	ユーザモデルの例	23
3.3	WWW サーバアクセスログの例	23
3.4	評価表	25
3.5	対称評価	27

Chapter 1

はじめに

今日、ネットワークの発展に伴い WWW や Netnews, mailing list といった新たな情報源が普及しつつある。我々はこうした情報源から多くの情報収集できるようになった。一方、収集した情報からユーザの必要とする情報を獲得するには多くの時間を要するようになった。このように情報収集の困難性が問題となっている。

これに対し新たな情報収集法として協調フィルタリングが存在する。協調フィルタリングは他者の知識を用いた情報収集法である。他者の知識とは、情報収集者の問題を解決する直接的な情報及び、問題解決に示唆を与える間接的な情報である。これらの情報は情報収集の時間と労力を削減する上で有用である。本章では、ネットワークからの情報収集の困難性、現在の情報収集法の問題点について述べる。

1.1 ネットワークからの情報収集の困難性

現在、ネットワークからの情報収集の困難性が問題となっている。問題の背景にはネットワーク上の情報源と流通する情報の増加がある。例えば、1996年1月の時点で国内のWWWサーバは12,204存在する[25]。更にこれらのWWWサーバ上には6,000万件を超えるWWWページが存在する¹。WWW以外にもNetnewsやMailing Listといった情報源が存在する。このため、情報収集者にとって情報源を特定して情報を獲得することが困難となっている。これを情報収集の困難性と呼ぶ。

情報収集の困難性に対して、いくつかの情報収集手法が提案されている。これらの手法を用いるユーザにとって次の点が問題となる。

- 求める情報が存在しないのに、あると信じて探し続けてしまう
- 求める情報が存在するのに、探し出せない/辿り着けない
- 不必要な情報ばかり収集してしまう

¹<http://www.goo.ne.jp/Help/faq7.html#a7>

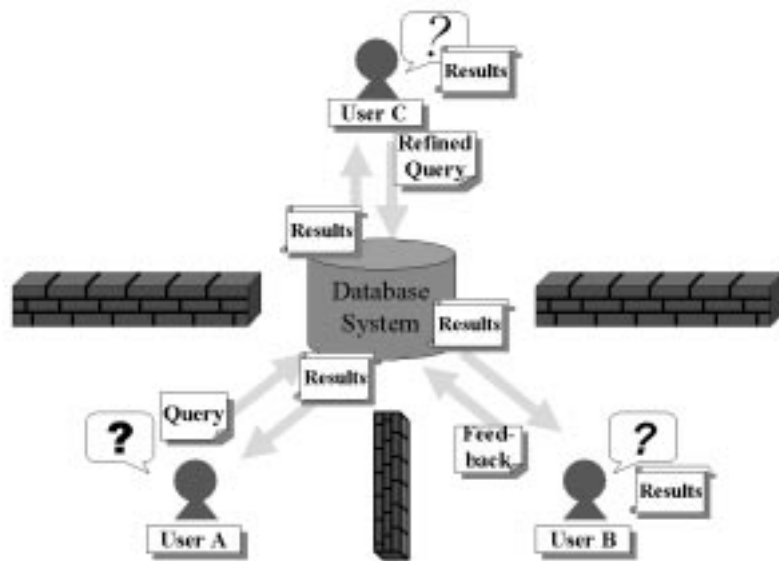


Figure 1.1: Information Retrieval

問題の背景には個々の情報収集者の熟練もさることながら、現在の手法にも一因がある。次節では現在の情報収集手法の問題について述べる。

1.2 現在の情報収集手法の問題

情報収集が独立して行なわれることが現在の情報収集手法の問題である。問題について述べる前に、現在用いられている情報収集手法について述べる。現在、ネットワークからの情報収集法として情報検索、情報フィルタリング、ブラウジングといった手法がある [24]。各手法について述べた後、問題点について議論する。

1.2.1 情報検索

情報検索とは、検索式の修正と検索結果に対するフィードバックの繰り返しによりユーザの求めるデータを特定していく情報収集法である。Figure 1.1 に情報検索のプロセスとユーザの関係を図示した。ユーザは求めるデータを表す検索式を立て、データベースに問い合わせる。データベースは検索式に合うデータをユーザに返戻する。ユーザは送られてきたデータ(検索結果)を評価する。ユーザの要求を満たす場合、ユーザは検索を終了する。検索結果がユーザの要求を満たさない場合、検索式を修正したり、検索結果の中から要求に近いデータを特定し(フィードバック)、再びデータベースに問い合わせる。情報検索は以上の繰り返しによりユーザの求めるデータを特定していく情報収集法である。

情報検索において、個々の検索過程及びユーザは独立である。このため、同じ関心を持

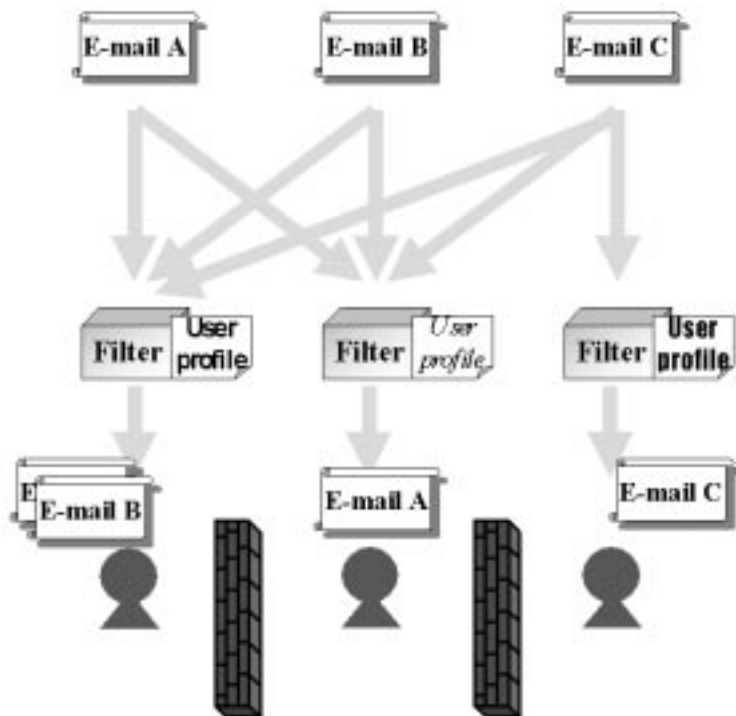


Figure 1.2: Information Filtering

ユーザーは互いに情報交換あるいは情報共有できない。効率的な検索は検索者個人の知識と熟練に委ねられている。

1.2.2 情報フィルタリング

情報フィルタリングは、ユーザーに対して送られてくる情報を除外したり、優先度を与える情報収集法である。Figure 1.2 に情報フィルタリングの概要を図示した。ユーザーに対し送られてくる情報の一例に Mailing List がある。参加者数の多い Mailing List では一日に 100 通以上のメールが流通する。こうした Mailing List に参加するユーザーにとって、e-mail の選別作業は時間のかかる作業である。情報フィルタリングは、この選別作業を自動化する。

情報フィルタリングの過程は次の通りである。ユーザーは除外するキーワードをプロフィールに事前に記述する。フィルタリングシステムはプロフィールに記述された情報を除外する。除外だけでなく、優先順位をつけることも可能である。プロフィールに優先順位をキーワードに与える。システムはプロフィールに従い、情報に優先順位をつけてユーザーに提供する。情報フィルタリングは情報を選別する情報収集法である。

情報検索と同様、個々のユーザーおよび過程は独立である。これによりユーザー間の情報共有は阻害され、効率的な情報収集は困難となる。各ユーザーは自身の知識に従いプロフィールを記述する。この記述により必要な情報まで除外してしまう恐れがある。このため、情

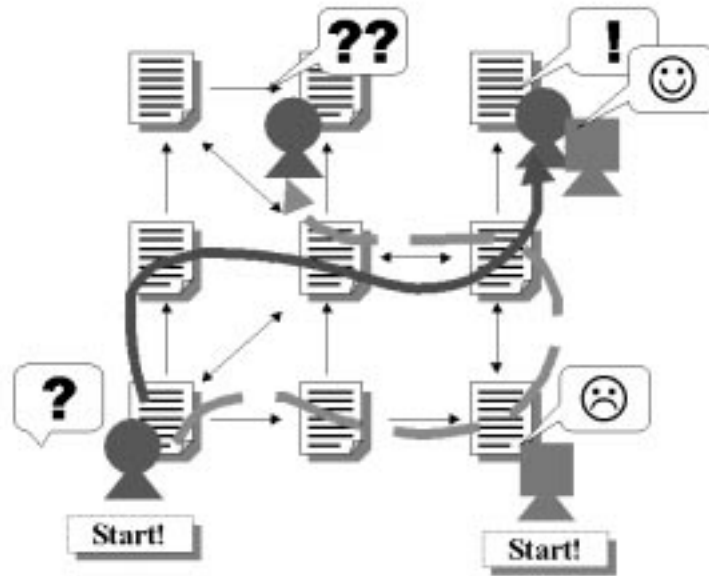


Figure 1.3: Browsing

報フィルタリングも情報検索と同様、効率的な情報収集には熟練を要する。

1.2.3 ブラウジング

ブラウジングとはハイパーテキストにおける情報収集法である。Figure 1.3 に略図を示す。具体例として WWW での情報収集が挙げられる。ブラウジングでは、ユーザの目的が明確でない場合がある。ユーザはブラウジングしながら目的を明確化し情報を獲得する。

ブラウジングは最も困難な情報収集法である。第 1 にユーザの必要とする情報が存在するか不明である。情報検索のように、全てのデータを対象に調査できないため、必要とする情報の存在は不明である。第 2 にユーザは必要とする情報にたどり着けない可能性がある。ユーザはリンクを辿りながら情報収集する。しかし、リンクの選択によっては必要とする情報と反対の方向に進んでしまう可能性がある。この問題は、ブラウジングが個々に独立していることに起因する。

ブラウジングにおいても個々のユーザあるいは情報収集過程が独立している。このため、ユーザは独力で情報収集せねばならない。

1.2.4 現在の情報収集手法の問題

情報収集過程およびユーザの独立が問題である。現在の手法では個々のユーザが独立している。これにより、ユーザ間での情報共有が阻害される。これは同じ関心を持った別の

ユーザが居たとしても互いに情報共有できないことを意味する。現実の状況では、同じ関心を持つユーザが居る場合、そのユーザと情報交換して効率的に情報収集を行なうことがある。しかし、現在の手法ではユーザは互いに独立しているため不可能である。効率的な情報収集にはユーザ間の協調や情報共有が必要である。

情報収集間に協調性を持たせることで効率的な情報収集が可能となる。協調的な情報収集の例として Mailing List がある。Mailing List では同じ関心を持った参加者が互いに情報交換を行なう。今、ある問題を持つユーザが参加者に問い合わせる場合を考える。Mailing List 参加者は質問に対して直接的な解答や、問題解決に関連する情報源や意見などを提示する。この過程を相互の協調による情報収集過程である。相互の協調により単独で情報収集するよりも効率的に情報収集できる。

ユーザ間の協調を支援する手法として協調フィルタリングがある。次章から協調フィルタリング技術の研究動向や課題について述べ、協調フィルタリングの可能性について探る。

Chapter 2

協調フィルタリングとは

協調フィルタリングとは、情報流通の仲介に基づき効率的な情報収集を目標とした情報収集手法である。情報流通の仲介とは以下を指す。

1. 個々の情報収集者が持つ問題解決のための情報を自動的に収集
2. 同じ問題を持つ情報収集者に提供

1. では、対象とする問題に関心を持つ者を探して、その者から問題解決のための情報を収集する。2. では、1. で得たユーザの特徴をに見合う情報を提供する。協調フィルタリングは、この過程を自動化し情報流通を仲介する。

現在、協調フィルタリングに対する期待が寄せられている。一例として、WWW ページの推薦¹ や Netnews のフィルタリング [14] をはじめ、音楽² や映画³ の推薦システム、ネットワーク上の有害情報を遮断するための Rating System や研究 / 教育のための情報収集手段 [34] などがある。今後、さらに多くの分野への応用が考えられる。

本節では協調フィルタリングについて概説する。協調フィルタリングの定義、関連概念、概念 / 定義の変遷、研究の経緯、これまでの研究について述べる。

2.1 定義

本稿における協調フィルタリングの定義を示す。

協調フィルタリング (Collaborative Filtering)

人間の情報収集行動から興味・関心・意図といった問題意識および獲得された情報を収集し、類似の問題意識を持った者に提供することで情報収集活動を支援するための情報収集の手法

但し、複数の人間が存在する環境を前提とする。これまでの研究においていくつかの定義が示されてきた。本稿ではそれらを含む包括的な定義を行なう。

¹<http://www.agentware.com/>

²<http://www.firefly.com/>

³<http://www-tate.iis.u-tokyo.ac.jp/tatemura/CinemaScape/>

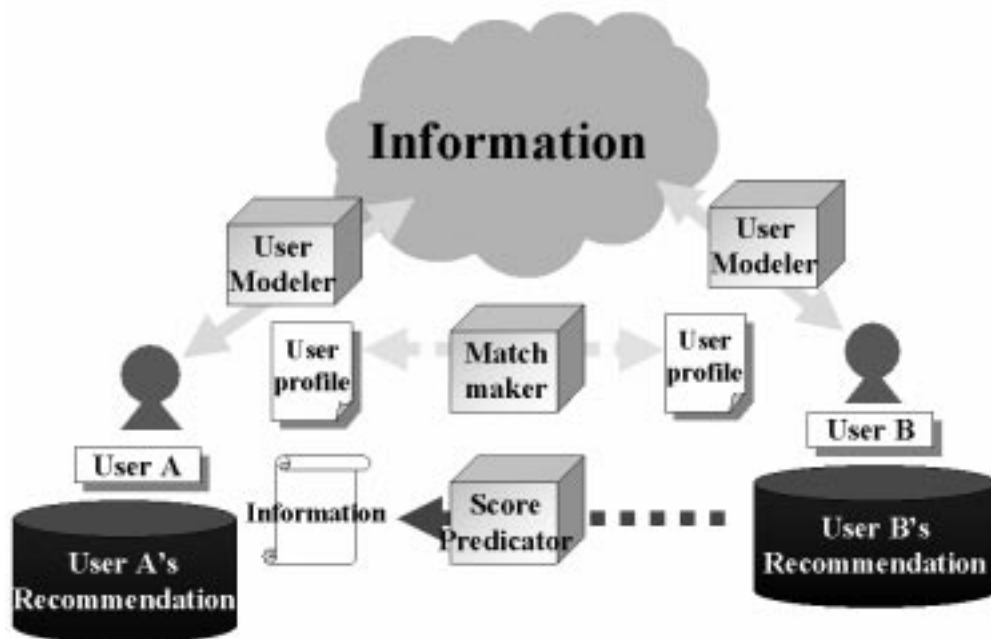


Figure 2.1: Collaborative Filtering

2.1.1 アーキテクチャ

協調フィルタリングシステムのアーキテクチャを示す。協調フィルタリングシステムは以下の要素から成り立つ。

User profile 情報の取捨選択の基準を記述するファイル

Recommendation ユーザが情報収集で得た情報を蓄積するファイル

User modeler ユーザの特徴を抽出する処理単位

Match maker ユーザ間の類似度を計算する処理単位

Score predictor 個々のユーザに応じて情報のランキングを計算する処理単位

Figure 2.1 に各要素の関係を示す。

User profile はユーザの関心を記述するためのファイルである。User profile には、ユーザの関心を表すキーワードや特定のデータ項目 (e-mail での From: フィールドなど) などが記述される。システムは User profile を元に情報の取捨選択を行なう。この他、ユーザ間の類似度計算にも用いられる。User Profile の例を Table 2.1 に示す。

Recommendation はユーザの情報収集行動を通じて情報を蓄積するファイルである。ユーザの関心に応じて情報収集した結果、得られた情報 (URL や e-mail address など) が蓄積

Table 2.1: User profile の例

Item	Score or Data
情報検索	+2.0
Agent System	+3.5
Field == Date	5/Mar
URL == http://www.media.mit.edu/	-1.5

Table 2.2: Recommendation の例

Recommendation
情報検索: Gerard Salton(http://simon.cs.cornell.edu/), Nicholas J. Belkin(http://sunsite.informatik.rwth-aachen.de/) Software Agent: Pattie Maes(Pattiemaes@media.mit.edu), Henry Leiberma(HLeibe@media.mit.edu)

される。これらの情報にはユーザの評価に応じて有用度が与えられる。Recommendation 中の情報は他者への情報提供に利用される。Recommendation の例を Table 2.2 に示す。

User modeler はユーザの行動から関心を獲得し User profile を更新する処理単位である。ユーザの関心は動的に変化する。さらにユーザ自身が関心をキーワードとして表せない場合も考えられる。このため User profile の記述を手動で行なっているのは非効率である。User modeler はユーザの情報収集行動から特徴を観察し、ユーザの関心を profile に記述する。

Match maker はユーザ間の関心の類似性を測定する処理単位である。具体的には評価式を用いて User profile 間の相関を計算する。協調フィルタリングでは、関心の類似するユーザを情報提供者として利用するため、類似する関心を持つユーザの特定を行なう必要がある。Match maker は、ユーザの関心を表す User profile を比較することにより、ユーザ間の関心の類似性を測定する。

Score predictor は個々のユーザに応じた情報の予想評価を計算する処理単位である。Score predictor は Recommendation に記述された有用度とユーザ間の類似度を用いてユーザ毎の評価を予想する。

なお、協調フィルタリングシステムの入力 / 出力は次の通りである。

入力 Recommendation, User profile

出力 個々のユーザに応じて評価された Recommendation

以下、シナリオに沿ってこれらの機能を説明する。ユーザは問題に関する情報を収集するため web をブラウジングする。User modeler はユーザの行動から特徴を見つけ、ユーザの関心を推測する。ユーザの関心を表す特徴として・ユーザがページを眺める時間、・各

ページに含まれる語句などが考えられる。User modeler はこれらの情報を元に User profile にユーザの関心を記述する。システムは協調フィルタリングを行なうため、同じ関心を持つユーザを特定する。このため、Match maker は作成した User profile と他のユーザの profile を比較し、関心の類似するユーザ (類似ユーザ) を特定する。ユーザは過去に獲得した情報を Recommendation として蓄積している。システムは類似ユーザを特定した後、Recommendation からユーザの関心に合う情報を探す。このため、Score predictor を用いてユーザの評価を予想する。Score predictor はユーザの類似度と Recommendation からユーザに提示する情報の評価を予想する。システムは評価予想が閾値を超える情報をユーザに提示する。以上が協調フィルタリングシステムのアーキテクチャである。

2.2 関連概念

協調フィルタリングの関連概念について示す。

Information Filtering

ユーザの関心に従って自動的に情報収集する技術。協調フィルタリングの上位概念。以下に分類できる。情報フィルタリングの分類については森田 [32]、Malone [9] の文献を参照されたい。

Content-based Filtering / Cognitive Filtering

キーワードと内容の比較によるフィルタリング。本稿では内容に基づくフィルタリングと呼ぶ。

Social (Information) Filtering

実社会の人間関係に基づくフィルタリング。例えば上司からのメールは優先させるなど。本稿では社会的なフィルタリングと呼ぶ。

Economic Filtering

情報の生産と消費のコストとメリットに基づくフィルタリング。例えば、生産が容易な情報 (Netnews におけるクロスポストなど) は生産コストが低いため得られるメリットが少ない、など。

情報検索 (Information Retrieval)

検索質問に従い情報を収集する技術。情報フィルタリングとの相違は情報収集過程にある。

Recommender System

情報を必要とする者に応じた情報を提供するシステム。本稿では推薦システムを用いる。

Rating-based Filtering

協調フィルタリングの一種。関心の表現に得点を用いる。本稿では得点に基づくフィルタリングを用いる。

Table 2.3: 情報検索と情報フィルタリング

	情報検索	情報フィルタリング
情報源	一定 / 静的	不定 / 動的
情報の流れ	双方向	一方向 (情報源 利用者)
関心	短期 / 動的	長期 / 静的

以上、協調フィルタリングに関連する概念について述べた。次に、類似の概念間の相違について述べる。

協調フィルタリングと推薦システムはほぼ同一の概念である。Resnick は両者の相違について次のように述べている。両者の相違は、・関心の類似するユーザの特定が自動的に行なわれるか否か、・情報を除外するだけでなくユーザの関心に応じた情報提供を行なうか否か、という2点にある [15]。推薦システムはユーザの特定を自動化し、ユーザの関心に応じた情報提供を行なう。協調フィルタリングはユーザ自身が関心の同じユーザを特定し、不必要な情報の除外を行なう。

本稿における協調フィルタリングは Resnick らの言う推薦システムと同義である。本稿では協調フィルタリングを、関心の類似するユーザを自動的に特定し、個々のユーザの関心に応じた情報を提供する情報収集手法と見なす。

次に情報フィルタリングと情報検索との相違について述べる。情報フィルタリングと情報検索は互いに類似する概念である。実際、情報フィルタリングは情報検索技術の応用である。以下では情報検索と情報フィルタリングの相違について述べる。

2.2.1 情報検索と情報フィルタリング

情報検索と情報フィルタリングの違いは、・情報源の性格、・関心の持続期間、・情報収集の方法にある。両者の相違点を Table 2.3 に示す。

以下、Table 2.3 について説明する。

情報源 情報源の特徴を指す。

- 情報検索における情報源は静的である。
情報検索から見た情報源は一つまたは複数のデータベースである。情報フィルタリングにおける情報源と異なり、ユーザが選択した情報源に対して検索を行なう。ユーザが変更するまで変わらない。以上から情報検索における情報源は静的である。
- 情報フィルタリングから見た情報源は動的である。
情報フィルタリングの対象とする情報源は複数存在する場合がある。メールや Netnews の記事、WWW ページといった様々な情報源があり、状況に応じて情

報源は変化する。情報検索の様に決まった情報源を常に利用することがない。この意味で情報フィルタリングにおける情報源は動的である。

情報の流れ ユーザと情報源の間の情報の流れる方向を指す。

- 情報検索での情報の流れは双方向である。
情報検索過程は検索と結果に対するフィードバックにより成り立つ。このため、情報の流れは双方向である。
- 情報フィルタリングにおける情報の流れは一方向である。
情報フィルタリングでは情報源からユーザに対して情報が流れる。このため、情報の流れは一方向である。

関心 ユーザの関心の持続期間を指す。

- 情報検索ではユーザの関心の持続期間は短期的である。
情報検索では、ユーザはその時の必要に応じて検索を行なう。ユーザは検索結果から検索質問を修正したり新たな検索質問を立てたりする。この意味でユーザの関心は短期的であり動的である。
- 情報フィルタリングにおける関心の持続期間は長期的である。
ユーザはプロフィールに必要とする情報や除外する情報を記述する。プロフィールは情報検索と異なり、数週間あるいは数カ月程度の長期的な関心が記述される。さらにプロフィールは頻繁な修正を想定していない。以上から、情報フィルタリングにおけるユーザの関心は長期的であり静的である。

情報検索と情報フィルタリングに関して参考となる文献を次に示す。情報フィルタリングと情報検索の相違については Belkin[2] の文献が有用である。情報検索研究については住田 [23] の文献が有用である。情報フィルタリング研究については森田 [32]、Oard[11]、Goldberg[5] の文献が有用である。

2.3 これまでの研究

これまでの協調フィルタリング研究について述べる。ここでは以下に示す研究について述べる。

- Tapestry
- GroupLens
- Ringo
- Fab
- PHOAKS

以下各研究について、概要、内容、研究課題を述べる。

2.3.1 Tapestry

概要

Tapestry[5] は e-mail, Netnews のフィルタリングを行なうシステムである。Tapestry は協調フィルタリングの概念を初めて示したシステムである。

内容

システムはキーワードを用いて情報をフィルタリングする他、同じ関心を持つユーザ名を指定してフィルタリングする。これは自分と関心の同じユーザは、自分と同じように情報を評価する、という仮定に基づいている。システムは e-mail や Netnews の記事とそれに対する評価 (リプライ、コメント) を蓄積する。ユーザからの要求に応じてシステムはこれらの評価をフィルタリングに利用する。

研究課題

研究課題としてプライバシーの保護、同じ関心を持つ他者の探索が挙げられる。システムは個人のメールと Netnews 記事を同一に扱うため、プライバシーの保護が必要である。また、ユーザは同じ関心を持つ他のユーザを事前に知っておく必要がある。これは一つの機関や組織といった限られた範囲であれば良いが、複数のネットワークで接続されたシステムを対象とする場合、同じ関心を持つ別のユーザを探すことは困難である。

2.3.2 GroupLens

概要

GroupLens[14] は Netnews を対象とした協調フィルタリングシステムである。ユーザの評価を得点で表すこと、集めた得点を元に個人の得点を予想することが特徴である。

内容

ユーザは記事を読み、それに対して 5 段階の得点を与える。システムは記事と各ユーザの得点を収集する収集した得点の集計結果を元に個々のユーザの得点を計算する。計算方法については 3.2にて述べる。

本研究と従来研究との比較が示されている。Tapestry との比較では、事前に同じ関心を持つユーザを知っておく必要がないことが挙げられている。また、フィルタリングを行なうサーバを分散させることで負荷の軽減を図った。この他、評価者の匿名性を挙げている。

研究課題

問題点として以下が挙げられる。

1. 誰も読んでいない記事は得点予想できない
2. 得点予想が正確になるまでに時間がかかる

システムはユーザからの得点という明示的なフィードバックを必要とする。このため、誰も読んでいない記事に対する得点予想は行なえない。更に、個々の得点予想の精度向上には多くの得点が必要となる。精度の高い得点予想を行なうには、記事に対する多くの得点が必要であり、多くのユーザを必要とする。このため、ユーザと得点の獲得に時間を要することが問題である。

2.3.3 Ringo

概要

Ringo[18] は音楽の推薦を行なう協調フィルタリングシステムである。GroupLens と同様、ユーザからの得点を元に個々のユーザに応じた得点予想を行なう。この研究では、得点予想の計算法について比較評価を行なった。

内容

評価予想の精度改善について比較調査した。Ringo ではピアソン積率相関係数をはじめ、いくつかの評価法について比較調査した。比較対象は、実際の得点と予想得点との誤差、集計した得点から予想得点を計算できた割合である。詳細については 3.3 を参照されたい。

研究課題

問題点は GroupLens の場合と同じである。ユーザに得点という明示的なフィードバックを求める以上、誰も見ていない記事は得点予想できない。また、ユーザに近い関心を持つ別のユーザを必要とする。このユーザが一定数に達しない限り、精度改善は見られない。

2.3.4 Fab

概要

Fab[1] 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双方の特徴を持つ推薦システムである。両者の特徴を持つことで、GroupLens や Ringo の問題を改善する。

内容

筆者らは従来の協調フィルタリングシステムの問題点として以下を示している。

1. 誰かが情報を評価しなければならない
2. 多くの評価を必要とする

これらは GroupLens, Ringo に共通する問題である。

1 点目は根本的な問題である GroupLens, Ringo と同様に、情報に対するユーザからの明示的な評価を元に予想を行なう。推薦する情報は、過去に誰かが評価している情報でなくてはならない。この問題は情報の増加とともに大きくなる。

2点目は精度と時間の問題である。精度の高い予想には、類似する関心を持ったユーザの評価を必要とする。システムの提供する予想は、関心の類似するユーザの評価を多く集めるほど正確になる。しかしながら、マイナーな関心を持つユーザに対しては、他の類似ユーザを見つけることが困難であることから、精度の高い予想は困難である。このため、精度の向上には、各々が多種多様な関心を持つユーザを多く獲得する必要がある。また、ユーザを獲得しても、そのユーザから評価を得なければならない。新たに出現した情報が評価を得るには時間を要する。このように、ユーザと評価の獲得に時間を要する。

Fab は協調フィルタリングと内容に基づくフィルタリングの構成によるハイブリッドなシステム構成を採用することにより上の問題を解決した。Fab システムはマルチエージェントシステムを採用している。システムは次のエージェントから構成される。

Collection Agent プロファイルに記述された主題についてネットワークから情報収集する

Selection Agent ユーザ毎に生成され、Collection Agent の収集した情報からユーザに応じた情報を取得する

central router Collection Agent と Selection Agent の仲介を行なう

システムの動作は次の通り。Selection Agent はユーザの関心に応じて Collection Agent から情報を取得しユーザに提供する。ユーザはこれに対して 7 段階の評価を返す。Selection Agent はユーザからの評価を Collection Agent に返す。Collection Agent は評価を元に自分のプロファイルを更新し、次回の情報収集に反映する。また、システムは定期的に Collection Agent を更新する。更新の目的は Collection Agent の収集する情報の精度向上である。更新により Collection Agent の削除・複製が行なわれる。更新は Collection Agent が Selection Agent に情報提供した度合やユーザからの評価に応じて行なわれる。システムは Collection Agent の削除と複製を繰り返しながらユーザに提供する情報の精度向上を目指す。

Fab は協調な情報フィルタリングである。従来の協調フィルタリングはユーザ相互の情報共有により成り立つのに対し、Fab はユーザからのフィードバックにより Collection Agent の性能向上を図る。個々のユーザのフィードバックによりシステム全体の性能を向上させ、結果的に個々のフィルタリングの精度向上を図る。このことから、Fab は協調的な情報フィルタリングである。

結果、時間とともにフィルタリングの精度向上が見られた。また、Collection Agent の更新により時間とともに、Collection Agent のプロファイルが専門的になることが観察された。

研究課題

システムがユーザに適合するまでの時間 (適合時間) が問題である。システムはユーザからの評価に応じて Collection Agent の持つプロファイルを更新し、定期的に

Collection Agent の削除・複製する。しかし、システムがユーザに適合するには時間がかかる。筆者らは適合時間について議論していない。さらに Collection Agent の更新方法についても議論する必要がある。

2.3.5 PHOAKS

概要

PHOAKS[20] は URL 推薦システムである。GroupLens や Ringo のような得点に基づくフィルタリングと異なり、PHOAKS は自動的に Netnews から推薦情報 (URL) を獲得する。

内容

自動的な推薦獲得がシステムの利点である。PHOAKS は Netnews からキーワードと URL を取得する。対象となる記事は URL が記載されている記事および FAQ である。URL 抽出の方針は以下の通りである。

1. クロスポストされた記事に記載された URL の除外
2. 署名に記載された URL の除外
3. 前の記事のリプライ部分に記載された URL の除外
4. テキストのブロックで囲まれた URL を取得

筆者らは課題として、推薦提供者と受信者間における信頼性および類似性の評価方法の確立を挙げている。推薦の質と精度の向上にはこの評価方法を検討する必要がある。

研究課題

推薦情報を既に存在する情報に依存するため新たな推薦を促進させない点が問題である。推薦の提供には、多くの推薦を獲得する仕組みが必要である。現在のシステムの推薦取得機構は、過去にポストされた記事から URL を獲得するため受動的である。より多くの推薦を獲得するために、積極的に推薦を引き出す必要がある。このため、Netnews の議論を促進させ、推薦情報が発生する機会を増やす機構が必要である。

Chapter 3

協調フィルタリングにおける研究対象

現在行なわれている協調フィルタリングの研究対象について述べる。主な研究対象はユーザの行動からの関心獲得と提供する情報の評価予想である。以下に研究対象を示す。

ユーザモデリング 明示的な情報からユーザの関心を獲得しモデル化

類似度計算 ユーザ間の類似性を計算

評価の予想 各ユーザに応じた評価の予想と精度の向上

以下、順に述べる。

3.1 ユーザモデリング

協調フィルタリングではユーザの関心をモデリングする。このユーザモデリングにより、システムはユーザに代わって情報を評価し自動的に情報収集する。

以下ではユーザモデリングに関するこれまでの研究について述べる。これまでの研究として情報検索分野、WWW 情報空間での行動分析の例を取り上げる。

3.1.1 情報検索におけるユーザモデリング

情報検索分野において、ユーザモデリングは個人に適応した検索システム構築のための重要な研究課題である。現在の検索システムはユーザを区別しない。一方、個々のユーザは各々異なる検索要求および領域知識を持つ。たとえ検索要求が同じであっても、求める情報の専門性や種類に違いが存在する。個人に応じて精度の高い検索を提供するため、システムはユーザの情報要求や領域知識を獲得せねばならない。このため検索システムはユーザモデリングを必要とする。

Ingwersen はユーザモデルの目的と定義を次のように述べている [6]。

ユーザモデル ユーザの検索要求を特定するためのモデルであり、ユーザの嗜好と価値、期待、志向性、情報検索知識、領域知識などを表す。



Figure 3.1: Oddy の連想グラフ (Ellis の文献より)

Ingwersen の述べるよう、個人に応じて検索を行なうシステムを構築するためにはユーザの嗜好を獲得するためのユーザモデルが必要である。

情報検索における研究には Oddy, Belkin, Rich らの研究がある。Ellis の文献 [4] は、Oddy や Belkin によるグラフを用いたモデル、Rich の典型を用いたモデルを示している。グラフを用いたモデルは扱いやすさと人間の理解の容易さから多くの研究で用いられている。典型を用いたモデルは、特徴ある典型を予め用意しておき、その組合せにより個々のユーザをモデリングする。以下、順に説明する。

グラフによるユーザモデリング

グラフ表記によるユーザモデル表現が多く行なわれている。こうした研究には Oddy [12], Belkin [3], Jennings [7] などがある。これらの研究で、基本的な考えは同じである。相違点はノードとノード間の重み付け方法である。以下では共通的なアイデアの示されている Oddy の研究について説明する。

Oddy は THOMAS システムで連想グラフを用いたユーザモデルを示した。図 3.1 にグラフを示す。ユーザモデルは文献、主題語、著者名をノードとしたグラフにより表される。図中で四角は文献ノードを、数字は文献番号を表している。文献ノードとリンクしている語はキーワードである。

グラフ作成は以下の手順で行なわれる。ユーザはシステムに検索要求を表すキーワードを示す。グラフ中にキーワードが存在しなければ、システムは新たなノードとしてキーワードを追加する。逆にキーワードがノードに存在する場合、そのノードを辿って文献ノードを特定し、文献ノードの所属度 (involvement) を計算してユーザに提示する。ユーザはシス

Table 3.1: Feminist の典型

属性	属性値	評点
Genres Woman	3	700
Piety	-5	800
Conflicts Upbringing	3	700
Strength Perspective	3	600

テムの提示した文献に適不適を示す。不適の場合、その文献ノードは連想グラフから除外される。以上の繰り返しによりグラフを修正していく。

この問題点としてグラフ構築の困難性が挙げられる。システムはグラフを拡張・修正するため、ユーザにキーワードを要求する。この時、システムの提供する文献がユーザにとって不適なものばかりだと、ユーザはその度キーワードを入力せねばならない。こうした作業は労である。このため、予めユーザの概念体系を包括するようなグラフ、またはユーザのキーワード入力の負担を軽減する機構が必要である。しかし、ユーザの関心を含む包括的なグラフは大規模であり、構築に多くの時間を要する。各ユーザの関心に応じてグラフを作成することはさらに困難である。連想グラフは人間にとって理解しやすいが、構築には困難を伴う。

典型によるユーザモデリング

Rich は典型を用いたユーザモデリングを示した [16]。Rich の作成した GRUNDY システムは、個々のユーザを予め用意された典型 (stereotype) の組合せでモデリングする。システムは個々のユーザモデルに応じた文献を推薦する。

典型には属性が設定されている。属性とは典型を特徴付けるキーワードである。各典型は属性と属性に対する得点 (-5 から +5)、評点を持つ。得点は典型と属性の関連性を表す。得点が高いほど高い関連性を示す。評点とは典型の強弱を示す数値である。Table 3.1 に典型の例を示す。

ユーザモデリングは次のように行なわれる。システムは典型の組み合わせによりユーザモデリングを行なう。システムはユーザの入力したキーワードから典型を選択する。典型の選択は次の場合に行なわれる。・ユーザの示すキーワードが典型中の属性に一致する場合、もしくは・直接ユーザが属性に値を与える場合、である。これにより典型の組合せが行なわれる。この結果、ユーザモデルが構成される。Table 3.2 に例を示す。

文献の提示は次のように行なわれる。組合せたユーザモデル中の属性値と評点が一定の閾値を超えた時、システムはユーザに文献を提示する。

典型を用いたユーザモデリングには利点と欠点が存在する。利点には、簡単な典型の組合せで複雑なユーザモデルを表現できる点がある。システムは個人に応じてモデルを構築する必要がない。欠点は上の逆で、完全に個人に適應できない点である。典型は特徴的なユーザのモデルである。システムは典型の組合せによりモデルを構築するが、あくまでユー

Table 3.2: ユーザモデルの例

属性	属性値	評点	理由 (組合せの元となった典型)
Piety	-3	423	Woman, Feminist, Intellectual
Thrill	-4	839	Woman, Intellectual
Motivations Learn	4	700	Intellectual
Weakness Reason	3	600	Intellectual

Table 3.3: WWW サーバアクセスログの例

host name	date	request	server code	bytes
24cpced.aist-nara.ac.jp	[02/Mar/1998:07:57:30 +0900]	"GET /ats/recent.gif HTTP/1.0"	200	229376
24cpced.aist-nara.ac.jp	[02/Mar/1998:07:57:33 +0900]	"GET /ats/weekly.gif HTTP/1.0"	200	196608
24cpced.aist-nara.ac.jp	[02/Mar/1998:07:57:41 +0900]	"GET /ats/ats.map?244,185 HTTP/1.0"	302	-
24cpced.aist-nara.ac.jp	[02/Mar/1998:07:57:43 +0900]	"GET /ats/country.gif HTTP/1.0"	200	327680
24cpced.aist-nara.ac.jp	[02/Mar/1998:07:58:56 +0900]	"GET /ats/country.gif HTTP/1.0"	304	-

ザの近似でしかない。完全にユーザに特化したモデルは新たに属性の追加や属性値の変更などを行なって、はじめから構築する必要がある。

上の問題はトレードオフである。問題の状況に応じて典型を用いるか決めねばならない。ユーザモデルに正確な精度を求めないのであれば組合せで足りる。正確な精度を求めるなら、別のモデリングを考慮する必要がある。

3.1.2 行動履歴に基づくユーザモデリング

情報収集における行動履歴から特徴を抽出することで行なうユーザモデリングを示す。例として WWW 情報空間におけるユーザの行動の追跡について述べる。近年の web の普及とともに、web 上での行動履歴をユーザモデリングとして利用する研究が進められている。こうした研究には、WWW サーバアクセスログの利用から、検索エンジンへの検索質問の利用、複数のブックマークを用いたモデリングなどが存在する。以下、各々について述べる。

WWW サーバアクセスログを用いたユーザモデリング

WWW サーバアクセスログを利用したユーザモデリングについて述べる。WWW サーバアクセスログとは、クライアントとサーバ間での対話を記録したものである。表 3.3 に例を示す。アクセスログには、送信したページや取得したホスト名、時刻などが記載される。こうした情報はユーザの関心を知る上で一つの手段である。

三浦らは NCSA の httpd におけるアクセスログからユーザを特定し、ユーザモデリン

グの可能性について述べた [28][29]。ログからユーザを特定¹し、各ユーザのページ取得履歴からユーザの行動の特徴について調査した。この結果、WWW ログから各ユーザ毎に応じた特徴を検出した。しかし、アクセスログだけではユーザモデリングが困難なことが示されている。何故ならばアクセスログは“どのホストから、いつ、どのページにアクセスがあった”といった程度の情報でしかない。ユーザのブラウジング時間やユーザの辿ってきたリンク履歴などはアクセスログから取得できない。このため、アクセスログだけではユーザモデリングに適用できない。

一方、アクセスログはサーバ訪問者の全体的な傾向を把握する上で役立つ。すなわちページのアクセス数を示すことで協調フィルタリングとして利用できる。このため、全体的な傾向をランキングつまり人気投票としてユーザに提示する方法が考えられる。こうした研究には WWW 検索ログからユーザのニーズの把握 [21] や、ランキング提示 [30] などがある。

アクセスログを他の情報と組み合わせることで個々のユーザをモデリングした例もある²。この例ではアクセスログに記載されている各ページから単語を抽出し、単語の特徴ベクトルによりユーザの特徴を表している [22]。このサービスでは Cookie³ を用いてブラウザを特定する。これによりアクセスログからブラウザを特定 (ユーザを特定) している。このように、単にアクセスログだけを利用するのではなく、ページに含まれる情報とを組み合わせることで、ユーザモデリングが可能である。

以上、まとめると次のようになる。アクセスログだけによるユーザモデリングは困難であり、ログと他の情報を組み合わせたモデリングが必要である。また、サーバのアクセスログから得られる情報は少ないため、ブラウザ側にもユーザの情報を引き出す機構が必要である。一方、アクセスログはサーバ全体としての傾向を知る点で有効である。ユーザにアクセスの全体傾向を示すことで、ランキングという簡単な協調フィルタリングが可能となる。

ブックマークを用いたユーザモデリング

ブラウザのブックマークファイルを元にユーザモデリングを行なう。ブックマークはユーザの関心を知る上で有効な情報源である。ユーザは関心のある URL をブックマークに登録し必要に応じて参照する。さらに、登録された URL はフォルダ毎に分類され、個々のフォルダにはその特徴を表す名称が付与される。ブックマークは個人により異なるため特徴が現れる。以上から、ブックマークをユーザモデリングの情報として利用することが考えられる。

現在、ブックマークは URL の情報源として利用されており、ユーザモデリングには用いられていない。Rucker らの Siteseer [17] や森らのブックマークエージェント [31] は、ブックマークを URL の情報源として扱っている。これらは複数のユーザのブックマークを互いに共有し、必要に応じて利用するシステムである。

¹NCSA の httpd にはブラウザを特定する cookie が存在しないため、ユーザ特定が問題として取り上げられている。

²<http://smartpage.watch.impress.co.jp/>

³ブラウザを識別するための機能

Table 3.4: 評価表

<i>message No.</i>	Ken	Lee	Meg	Nan
1	1	4	2	2
2	5	2	4	4
3			3	
4	2	5		5
5	4	1		1
6	?	2	5	?

しかし、ブックマークを情報共有に用いるのではなく、ブックマークから特徴を抽出することが必要である。今後の研究として、ブックマークからのユーザモデリングが考えられる。

3.2 類似度計算

協調フィルタリングシステムではユーザ間の類似性を調べる必要がある。これは、次の理由による。

1. ユーザの評価を予想するため
2. システムに登録されたユーザから類似するユーザを特定するため

a

第1に、ユーザ間の類似度計算はユーザの評価を予想するために必要である。ユーザの評価の予想は得点に基づくフィルタリングにて研究されてきた。こうした研究には GroupLens や Ringo がある。これらはユーザの評価を得点として表し、これらの評価を元に個々のユーザの評価を予想する。

第2に、類似ユーザを自動的に発見するために必要である。協調フィルタリングでは、関心の同じユーザを特定する必要がある。これをユーザ自身が行なっているのは非効率である。このため、システムに登録されているユーザから類似ユーザを自動化する必要がある。

以下では、得点に基づくフィルタリングを例に GroupLens[14] の類似度計算と評価予想の計算法について述べる。

3.2.1 GroupLens における類似度計算

GroupLens はユーザから得た評価を表 3.4 のような配列として扱う。

まず類似度計算の手順について述べる。例えば、message 6 の Ken の評価について予想する。まず始めに Ken と同じ message を評価している他者 (この場合 Lee と Nan) の相関

係数 [13] を計算する。これを類似度とする。例えば Ken と Lee の類似度 r_{KL} は以下のよう
に算出される。

$$r_{KL} = \frac{Cov(K, L)}{\sigma_K \sigma_L} \quad (3.1)$$

$$= \frac{\sum_i (K_i - \bar{K})(L_i - \bar{L})}{\sqrt{\sum_i (K_i - \bar{K})^2} \sqrt{\sum_i (L_i - \bar{L})^2}} \quad (3.2)$$

$$= \frac{-2 - 2 - 2 - 2}{\sqrt{10} \sqrt{10}} = -0.8 \quad (3.3)$$

ここで \bar{K} は Ken の評価の平均、 \bar{L} は Lee の評価の平均を表す⁴ K_i, L_i は各々、message i に対する Ken と Lee の評価を表す。 σ_K, σ_L は各々 Ken と Lee の評価の標準偏差を表す。類似度は両者が評価したものについてのみ考える。同様に Ken と Meg の類似度は +1、Ken と Nan の類似度は 0 となる。つまり Ken の関心は Lee とは対称的で、逆に Meg の関心に類似する。一方、Ken と Nan の関心には相関がない。

次にユーザ間の類似度を用いて、個々に応じた評価を予想する。ここでは message 6 に関する Ken の評価 K_6 を予想する。評価の予想には次の式を用いる。

$$K_6 = \bar{K} + \frac{\sum_{J \in raters} (J_6 - \bar{J}) r_{KJ}}{\sum_J |r_{KJ}|} \quad (3.4)$$

$$= 3 + \frac{2r_{KM} - r_{KL}}{|r_{KM}| + |r_{KL}|} = 3 + \frac{2 - (-0.8)}{|1| + |-0.8|} = 4.56 \quad (3.5)$$

raters は Ken の評価に相関が存在する人の集合であり、この場合 Lee と Meg が相当する。 \bar{J} は各 *raters* の評価の平均を表す。⁵ J_6 は message 6 に対する *raters* の評価である。 r_{KJ} は Ken と *raters* の類似度 (相関係数) である。計算により Ken の message 6 に対する評価は 4.56 と予想される。同様に Nan の message 6 に対する評価も 3.75 と推定される。

上記で用いた予想方法は最小 2 乗誤差法であるが、ほかにも評価を予想する方法がある。Shardanand らは、最小 2 乗誤差法を含むいくつかの予想方法について比較評価している。

3.3 予想評価の精度

以下では、予想評価の精度を議論した文献について述べる。こうした研究に Shardanand らの Ringo が挙げられる [18]。この研究では、予想評価計算の 4 つの手法について予想評価の精度、予測率について議論している。

3.3.1 Ringo における予想評価の精度

Shardanand らは、複数の予想方法について精度と計算できる割合の比較を示した。この中で用いた方法は以下の通り。

⁴この例では簡単のため \bar{K}, \bar{L} に 3 を用いており、実際は異なる。

⁵ \bar{M}, \bar{L} に 3 を用いており実際は異なる。

Table 3.5: 対称評価

Plot A	User X	User Y
Artist 1	2	2
Artist 2	6	6
$r_A = +1.0$		
Plot B	User X	User Y
Artist 1	1	5
Artist 2	2	7
$r_B = +1.0$		
Plot C	User X	User Y
Artist 1	5	2
Artist 2	7	3
$r_C = +1.0$		

- 平均
- 最小 2 乗誤差法 (MSD:Mean Squared Difference)
- ピアソン積率相関係数
- 制約つきピアソン積率相関係数
- アイテム間相関アルゴリズム (Artist-Artist Algorithm)

これらについて、予想した評価とユーザの示した実際の評価の間の平均誤差と標準偏差、実際に計算できた割合 (予測率) を比較を行なった。

比較の結果、ピアソン積率相関係数に制約を加えたものが良い精度を示した。最小 2 乗誤差法を用いた予想では閾値 2.0 の時、平均誤差 1.2, 標準偏差 1.6, 予測率 70% を得た。一方、筆者らの示した制約付きピアソン積率相関係数を用いたものでは閾値 0.6 の時、上の同じ値で予測率 91% を得た。これにより、先行研究である GroupLens の精度を改善したと述べられている。

筆者らの方法は対称評価を用いている。対称評価とは互いに対称的な評価が行なわれることを評価を指す。例を表 3.5 に示す。表中にて、ユーザ X, Y は互いに対称的な評価を与えた。各 r はピアソン積率相関係数を表す。どの係数も全て +1.0 になっていることがわかる。つまりピアソン積率相関では、互いに対称的な評価にも関わらず同じ係数を生じさせている。

制約つきピアソン積率相関係数の計算法を以下に示す。

1. 対称評価を考慮したピアソン積率相関係数

$$\beta_{xy} = \frac{\sum(U_x - 4)(U_y - 4)}{\sqrt{\sum(U_x - 4)^2 \times \sum(U_y - 4)^2}} \quad (3.6)$$

β_{xy} を x, y の相関係数とする

2. 次に手順を示す。

- (a) ユーザ i と u の相関係数 β_{iu} を計算する
- (b) 閾値 L より大きい相関係数 β_{iu} を集め、これを類似集合 \mathcal{N}_u とする
- (c) 類似集合 \mathcal{N}_u の各ユーザ u について重み w_{iu} を計算する。

$$w_{iu} = \left(\frac{\beta_{iu} - L}{1 - |L|} \right)^2 \quad (3.7)$$

- (d) アーティスト j に対する全類似ユーザの重み付け評価の平均を評価予想 p_{ij} とする

$$p_{ij} = \bar{U}_i + \frac{\sum_k^{N_u} w_{ik} \times s_{kj}}{\sum_k^{N_u} w_{ik} \times c_{kj}} \quad (3.8)$$

最終的に比較を行なった (図 3.2)。 $|E|$ は平均誤差を、 τ は予測率を、 σ は標準偏差を表す。問題としているのは Extremes というユーザの好き嫌いが現れる値を取り出した部分である。

Method	All		Extremes		%T
	E	σ	E	σ	
Base Case	1.3	1.6	1.8	2.0	90
MSD, $L = 2.0$	1.0	1.3	1.2	1.6	70
Pearson r	1.1	1.4	1.5	1.7	99
Pearson r , $L = 0.35$	1.0	1.3	1.4	1.6	99
Pearson r , $L = 0.5$	1.0	1.3	1.3	1.6	95
Pearson r , $L = 0.65$	1.1	1.4	1.3	1.6	73
Pearson r , $L = 0.75$	1.1	1.5	1.3	1.7	41
Constrained Pearson r , $L = 0.5$	1.1	1.3	1.3	1.6	97
Constrained Pearson r , $L = 0.6$	1.1	1.4	1.2	1.6	91
Constrained Pearson r , $L = 0.7$	1.1	1.3	1.3	1.6	70
Artist Art st, $L = 0.6$	1.1	1.4	1.3	1.6	89
Artist Art st, $L = 0.7$	1.1	1.4	1.1	1.5	65

Figure 3.2: Summary of Results

筆者らが示した制約つきピアソン積率相関係数 (Constrained Pearson $r, L = 0.6$) は、Extremes において GroupLens(最小 2 乗誤差法) と同等であり、予測率で GroupLens を上回っている。

Chapter 4

今後の研究課題

協調フィルタリング研究の今後の課題として以下が挙げられる。

ユーザモデリング 暗示的な情報からの関心獲得

機械学習 ユーザの行動に応じてユーザモデルを動的に修正

マルチエージェントシステム 個々に自律した処理単位 (エージェント) により分散・統合型システムを構築

プライバシーの保護 ユーザの情報が他に洩れない機構を作成

4.1 ユーザモデリング

暗示的な情報を用いたユーザモデリングが必要である。暗示的な情報とは、ユーザが言明しないユーザの行動に現れる特徴である。こうした特徴には時間、ブラウジング経路などがある。ユーザはこうした情報をシステムに言明しない。これに対し得点は明示的な評価である。得点を用いることにより次の問題が存在する。

1. ユーザの負荷が大きい
2. 得点が偏ってしまう

第1にユーザに対する負荷が大きいという問題がある。これは得点が明示的な評価だからである。ユーザにとってはいちいち評価を与えねばならず負荷が大きい。

第2に得点が偏ってしまう問題がある。明示的な評価を得るにはユーザが関心を持つ必要がある。こうした情報は多くのユーザが関心を持つ一般的な話題である。このため明示的な評価を用いる場合、一般的な情報については得点を集めやすいが、専門的な情報については得点を集めにくいという欠点がある。このため、一般的な情報に得点が偏ってしまう問題がある。

問題解決には明示的な評価だけでなく暗示的な評価についても考慮したユーザモデリングが必要である。ユーザの行動という暗示的な情報からユーザの関心を獲得する技術が必要である。

暗示的な情報を用いたモデリングではデータマイニング (Data Mining) の応用が考えられる。データマイニングでは観察された事象から因果関係を発見する。ユーザモデリングにおいてはユーザの情報収集行動を観察し、その中からユーザの関心を表す行動を特定する必要がある。ユーザモデリングでは、ユーザの行動を収集し膨大なデータから関心を特定するため、データマイニングの応用を考える必要がある。

4.2 機械学習

ユーザモデリングではユーザに適応してプロファイルを作成する必要がある。このため、機械学習によるユーザへの適応が考えられる。ユーザの関心は常に変化するため、システムはユーザモデルを動的に修正する必要がある。平嶋ら [27] や畑島ら [26] は関心の減衰を用いたモデルを示した。これらのモデルにユーザからのフィードバックを合わせて関心の変化に対応することが考えられる。ユーザの変化する関心に適応するため、学習によるアプローチが考えられる。

現在、機械学習や関連概念を用いたユーザの行動予測が研究されている。Maes らの Learning Interface Agents [8] や Sheth の Newt システム [19]、Moukas のシステム [10] である。Maes のシステムはユーザの行動を観察し、システムが次に同じような状況に遭遇した際、ユーザに対して処理や助言を行なう。学習はユーザからの例示やフィードバックによって行なわれる。この結果、時間とともに行動予測が確実になったと述べている。Sheth はユーザからのフィードバックと遺伝的アルゴリズムを用いた情報フィルタリングシステムを構築した。システムはユーザからのフィードバックによりユーザに適応し、遺伝により新たな主題を獲得する。結果はユーザに適応したフィルタリングシステムを構築したと述べている。Moukas はマルチエージェントによる情報フィルタリングシステムを構築した。Moukas は各エージェントに競合や協調などの人工生命の概念を採り入れることでユーザへの適応を試みた。実験ではユーザの関心を静的な場合と動的な場合とに分け、システムのユーザへの適応を調査した。この結果、いずれの場合も時間とともにユーザへの適応が示された。

ユーザへの適応には機械学習によるアプローチが有効である。Moukas の例では機械学習がユーザの動的な関心に適応できることが示された。今後は適応時間の短縮が問題となる。このため、ユーザモデリングに複数の機械学習法を適用し、各学習法の適応時間の比較する必要がある。これにより最適の学習法を調査する必要がある。

4.3 マルチエージェントシステム

増加するユーザと情報への対応として、マルチエージェントシステムの応用が考えられる。協調フィルタリングシステムは、ネットワーク上に分散して存在するユーザの行動を

観察し、個別にモデリングしなければならない。これを集中型サーバ・クライアントシステムで行なうのは不適當である。分散した対象を扱うには、対象毎に処理単位を割り当て、処理単位間の協調によりシステムを構成することが望ましい。マルチエージェントシステムは個々の処理単位(エージェント)より構成される分散型情報システムである。協調フィルタリングシステムはネットワーク上に分散するユーザを対象とするため、マルチエージェント技術の応用が考えられる。

現在、マルチエージェントによる協調フィルタリングが試みられている。こうした研究に Fab[1] や宮原らのシステム [34] がある。Fab は・特定主題についてネットワークから情報収集する Collection Agent、・ユーザの関心に基づき Collection Agent から情報を選択する Selection Agent により構成される。Fab はこれらのエージェントの協調により情報収集、フィルタリングを行なう。宮原のシステムは個々のユーザモデルを持ったエージェントが自律的に情報収集を行なうマルチエージェントシステムである。システムは同じ関心によるグループにより情報共有を支援する。システムは・個々のユーザに対応してユーザモデリングするユーザエージェント、・グループでの情報共有を行なうグループエージェントにより構成される。

マルチエージェントによるシステム構築では、エージェント間の協調が問題である。マルチエージェントシステムでは、大規模なシステム構築に伴いエージェントの増加が見込まれる。これにより個々の交渉が増加するが、システムのアーキテクチャによっては破綻を来す可能性がある。個々の交渉が難航することで、結果として全体の性能低下を引き起こす可能性がある。このため、エージェントの増加にも対応できるアーキテクチャを考える必要がある。エージェント間の協調を効率化することはシステム全体の効率化にもつながる。このため、協調作業の効率化について考慮する必要がある。

さらに情報取引による交渉も問題となる可能性がある。協調フィルタリングは情報の経済活動と捉えることができる。現在の協調フィルタリングシステムは無償の情報提供により成り立っている。しかし、情報獲得さらに作成にはコストがかかるため、それに対する対価を求めることが考えられる。つまり、情報獲得に何らかの代償要求が考えられる。このため情報獲得と対価の折り合いの付け方が問題となる。MIT Media Lab. ではマルチエージェントによる経済活動の実験を行なっている¹。この実験では本や CD の売買をエージェント間の交渉により仲介する。エージェントは他のエージェントとの交渉により、ユーザの指定した物件をユーザの提示した価格で探したり売却する。エージェントは他のエージェントの関心に応じて価格の変更を行なう。協調フィルタリングシステムにおいても、指定された価格で情報を仕入れてきたり、値下げ交渉を行なったり、より高い価格で情報を提供したりするようなモデルが考えられる。このため、マルチエージェントシステムを単にシステム構築の側面から見るのではなく、エージェント間交渉の側面から捉えることも必要である。

¹<http://kasbah.media.mit.edu/>

4.4 プライバシーの保護

協調フィルタリングではユーザのプライバシーに関する情報を扱うため、セキュリティに注意する必要がある。システムはユーザの行動を観察し特徴を掴む。こうした情報は個人の興味や関心であり、個人が認めない限り外部に流出させてはならない。協調フィルタリングシステムはユーザプロフィールを厳重に管理しなければならない。

現在、ネットワーク上のプライバシーに関する指針が示されている。W3C² の P3P³ グループでは、個人および web サイトにおけるプライバシーの表現と取り扱いについての検討している。現在 P3P で検討されているドラフトに OPS⁴ がある。OPS はネットワーク上を流通する個人情報の取り扱いや表現を規定する。フィルタリングシステムを提供するベンダ^{5 6}は OPS を採用している。

協調フィルタリングシステムではプライバシーの保護機構を採り入れる必要がある。どこまでユーザをモデル化するか、プロフィールを誰と比較して良いかなど、ユーザの設定に従ってシステムは個人情報を取り扱わねばならない。マルチエージェントシステムでは、エージェント相互の交渉においてユーザのプライバシーを保護せねばならない。プライバシー保護は協調フィルタリングシステムにとって重要な課題である。

²World Wide Web Consortium

³P3P; Platform for Privacy Preferences Project
<http://www.w3c.org/P3P/>

⁴OPS; Open Profiling Specification

⁵<http://www.firefly.com/>

⁶<http://www.netperceptions.com/>

Chapter 5

おわりに

本稿では協調フィルタリングに関する研究動向について述べた。協調フィルタリングを概説し、現在の研究対象、今後の研究課題について述べた。

今後の研究課題でもユーザモデリングは重要な課題である。これまでの協調フィルタリング研究は、ユーザの評価予想に力点が置かれてきた。今後は暗示的な情報からユーザの関心を獲得し、ユーザの関心を学習するシステムが必要である。このため、データマイニング、機械学習が研究課題として挙げられる。また、プライバシー保護やマルチエージェントシステムの応用について述べた。

Bibliography

- [1] Balabanović, Marko., Shoham, Yoav. Fab: Content-Based, Collaborative Recommendation. CACM. Vol. 40, No. 3(1997)
- [2] Belkin, N. and Croft, B. Information Filtering and Information Retrieval:Two Sides of the Same Coin?. CACM. Vol. 35, No. 12, pp.29-38(1992)
- [3] Belkin, N. J. Anomalous states of knowledge as the basis for information retrieval. Canadian journal of information science, 5, 1980, 133-143.
- [4] Ellis, David.(細野 公男監訳) 情報検索論. 東京. 丸善株式会社, 1994, 180p.
- [5] Goldberg, David., Nichols, David., Oki, Brian M., Terry, Douglas. Using Collaborative Filtering to Weave an Information Tapestry. CACM. Vol. 35, No. 12(1992)
- [6] Ingwersen, Peter.(藤原 鎮男監訳) 情報検索研究：認知的アプローチ. 東京. トッパン, 1995, 378p.
- [7] Jennings, Andrew., Higuchi, Hideyuki. A Personal News Service Based on a User Model Neural Network. IEICE Trans. Inf. & Syst., Vol. E75-D, No. 2(1992)
- [8] Maes, Pattie., Kozierok, Robyn. Learning Interface Agents. Proceedings of the Eleventh National Conference on Artificial Intelligence, pp.459-465(1993)
- [9] Malone, T.W., Grant, K.R., Turbak, F.A., Brobst, S.A., Cohen, M.D. Intelligent Information-Sharing System. CACM. Vol. 30, No. 5. pp.390-402(1987)
- [10] Moukas, Alexandros. Amalthea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem. Proceedings of Conference on Practical Application of Intelligent Agents & Multi-Agent Technologies(1996)
- [11] Oard, D. W., Marchionini, G. A Conceptual Framework for Text Filtering. To be published in Journal of User Modeling and User-Adapted Interaction in 1997(1997) URL:<http://www.clis.umd.edu/dlrg/filter/papers/filter.ps>
- [12] Oddy, R. N. Information Retrieval through man-machine dialogue. Journal of documentation, 33, 1977, 1-14.
- [13] Pindyck, R.S. et.al. Econometric Models and Economic Forecasts. McGraw-Hill, NewYork(1991)
- [14] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. CSCW '94 Proceedings. pp.175-186(1994)

- [15] Resnick, P., Varian, H. R. Recommender Systems. CACM. Vol. 40. No. 3(1997)
- [16] Rich, E. A. User modelling via stereotypes. Cognitive science, 3, 1979, 329-354.
- [17] Rucker, James., Polanco, Marcos J. Siteeuser: Personalized Navigation for the Web. CACM, Vol. 40, No. 3(1997)
- [18] Shardanand, U. and Pattie Maes Social Information Filtering: Algorithm for Automating “Word of Mouth”. CHI '95 Conference Proceedings. pp.210-217(1995)
- [19] Sheth, Beerud Dilip. A Learning Approach to Personalized Information Filtering. Cambridge, MIT, 1994, 75p. MS. Thesis.
- [20] Terveen, Loren., Hill, Will., Amento, Brian., McDonald, David. and Creter Josh. PHOAKS: A System for Sharing Recommendations. CACM. Vol. 40, No. 3(1997)
- [21] 大久保ほか. WWW 検索ログに基づく情報ニーズ傾向の把握. 情報処理学会第 55 回全国大会, 2AC-5(1997)
- [22] 橘高 博行ほか. 個人適応型情報フィルタリングシステムの適用評価について. 情報処理学会第 55 回全国大会, 5Q-5(1997)
- [23] 住田 一男, 三池 誠司. 知的情報検索の動向. 人工知能学会誌. Vol. 11, No. 1(1996)
- [24] 武田 英明. ネットワークを利用した知的情報統合. 人工知能学会誌. Vol. 11, No. 5, pp.680-688(1997)
- [25] 中川 格, 石塚 英弘, 山本 毅雄. 日本の World Wide Web 情報空間: 1996 年 1 月のリンクデータ解析. デジタル図書館. Vol. 9(1997)
- [26] 畑島 隆, 元田 敏浩. 時系列情報を考慮したアクセスログ解析. 情報処理学会第 54 回全国大会, 4S-9(1997)
- [27] 平嶋 宗ほか. ブラウジングにおける動的興味の推定手法. 情報処理学会第 54 回全国大会, 4K-13(1997)
- [28] 三浦 信幸ほか. WWW サーバアクセス履歴からのユーザモデルの構築. 情報処理学会第 52 回全国大会, 4Aa-3(1996)
- [29] 三浦 信幸ほか. WWW サーバ・ログからのユーザモデル構築実験. 情報処理学会第 53 回全国大会, 3Q-2(1996)
- [30] 三浦 信幸ほか. WWW における検索ランキングとその Social Filtering としての効果. 情報処理学会第 55 回全国大会, 5Q-4(1997)
- [31] 森 幹彦, 山田 誠二. ブックマークエージェントによる URL の協調的情報フィルタリング. 人工知能学会誌第 29 回人工知能基礎論研究会 (1997)
- [32] 森田 昌宏. 情報フィルタリングに関する研究動向. JAIST Research Report. IS-RR-93-9I(1993). URL:<http://shinoda-www.jaist.ac.jp:8000/members/1993/>
- [33] 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- 個人の興味領域の表現およびその獲得法 - 信学技法. ET96-94(1996-12)

- [34] 宮原 一弘, 岡本 敏雄. 協調フィルタリングシステムにおけるエージェントの構成. 信学技法. AI97-3(1997-05)

Appendix A

WWW 上の情報源

協調フィルタリング研究に関する情報が WWW 上に存在する。多くは情報フィルタリングの一部として掲載されている。以下、ジャンル毎に URL とコメントを記す。

概要

Information Filtering Refs

情報フィルタリングに関する概要

<http://www.cs.kun.nl/is/research/filter/references.html>

Information & Collaborative Filtering

情報フィルタリングと協調フィルタリングとの相違について

<http://daniel.www.media.mit.edu/people/daniel/737/index.html>

リンク

Oard's Information Filtering Resources

情報フィルタリング研究についての大規模なリンク集。フリーの情報フィルタリングシステム、製品、研究論文、研究者へのリンクなど

<http://www.ee.umd.edu/medlab/filter/>

Berkeley Collaborative Filtering

1996 Collaborative Filtering workshop の案内をはじめ、協調フィルタリングシステム、研究者へのリンク、メーリングリストなど

<http://www.sims.berkeley.edu/resources/collab/>

Community-Based Navigation

Community-based Navigation という協調フィルタリング初期の研究の紹介

<http://community.bellcore.com/navigation/home-page.html>

プロジェクト / 研究論文

Software Agents Group

MIT Media Lab. Software Agents Group 研究テーマ

<http://agents.www.media.mit.edu:80/groups/agents/projects/>

GroupLens

GroupLens プロジェクト

<http://www.cs.umn.edu/Research/GroupLens/>

FAQ Finder

FAQ Finder プロジェクト

<http://infolab.cs.uchicago.edu/faqfinder/>

RefferalWeb

RefferalWeb プロジェクト

<http://www.research.att.com/~kautz/referralweb/index.html>

<http://infolab.cs.uchicago.edu/faqfinder/>

PHOAKS: People Helping One Another Know Stuff

PHOAKS プロジェクト

<http://www.phoaks.com/>

Social Information Filtering for Music recommendation

Ringo に関する論文

<http://lcs.www.media.mit.edu/groups/agents/papers/ringo/shard-thesis/ringo.ps>

ACM SIGCHI

ACM SIGCHI provides an international, interdisciplinary forum for the exchange of ideas about the field of human-computer interaction(HCI).

<http://www.acm.org/sigchi/>

研究課題

User modelling for Informatino Filtering

情報フィルタリングに関するユーザモデリング

<http://www.cs.su.oz.au/~bob/um96-workshop.html>

P3P; Platform for Privacy Preferences Project

個人情報の取り扱いに関するワークグループ

<http://www.w3.org/P3P/>

OPS; Open Profiling Specification

W3C に提出された個人情報取り扱いのためのドラフト (Firefly, Netscape Communications, Microsoft による)

<http://www.w3.org/Submission/1997/6/>

PICS; Plat for Internet Content Selection

Rating system における rating 情報取り扱いに関するワークグループ

<http://www.w3.org/PICS/>

研究者

Pattie Maes

Associate Professor at MIT Media Lab. Software Agents Group.

<http://lcs.www.media.mit.edu/people/pattie/>

Alexandros Moukas

PhD student at the MIT Media Lab. Software Agents Group. His interests include E-Commerce, Adaptive/Distributed Multiagent Systems and Artificial Life.

Amalthea 開発者

<http://moux.www.media.mit.edu/people/moux/>

Alan Wexelblat

Ph.D. student at the MIT Media Lab.

His doctoral research centers around the notion of using history information to help people solve present problems.

<http://wex.www.media.mit.edu/people/wex/>

Paul Resnick

Associate Professor at University of Michigan.

GroupLens 開発者

<http://www.si.umich.edu/~presnick/>

Henry Kautz

technical manager for the artificial intelligence area at AT&T Labs.

RefferalWeb, HiddenWeb プロジェクト

<http://www.research.att.com/~kautz/>

Loren Terveen

His research interest is computer-mediated communication and collaboration. Specifically, He's working on recommender systems.

PHOAKS 開発者

<http://www.research.att.com/~terveen/>

Marco Balabanovic

PhD student at Stanford University Robotics Lab.

Fab 開発者

<http://robotics.stanford.edu/people/marko/>

Doug Oard

Assistant Professor in the College of Library and Information Services at the University of Maryland

His research interests include cross-language text retrieval, information filtering, and speech-based information retrieval.

<http://www.glue.umd.edu/~oard/Welcome.html>

その他

これ以外については下記を参照されたい。

情報フィルタリング研究リンク集

http://cairo.aist-nara.ac.jp/~tomohi-f/jif_link.html