

4章 推薦アルゴリズムの概要

2023年1月14日 13:05

- 本章では、推薦システムの定義である「複数の候補から価値のあるものを選び出し、意思決定を支援するシステム」の前半部分、「複数の候補から価値のあるものを選び出す」ことを実現する推薦システムのアルゴリズムについて説明
- 具体的には、サービス内の大量のアイテムの中からユーザーに推薦するアイテムを選ぶために、システムに蓄積されたユーザーの嗜好情報やアイテムのコンテンツ情報などさまざまなデータに基づき、ユーザーがどのアイテムを好むのかを計算するアルゴリズムの説明
- 1章で説明した推薦システムの3つの構成要素の中の「プロセス（推薦の計算）」に該当
- 推荐システムのアルゴリズムには、機械学習を利用するような高度な手法もあれば、経験に培われた人手によるルールベースに基づくものまでさまざま
- それぞれが異なる特徴を備えており、常にこれを使えば良いというものはない
- 実際に推薦システムを開発するにあたって適切なアルゴリズムを採用するためには、それぞれのアルゴリズムの特徴を押さえた上で状況に応じたものを見極める必要がある
- 本章では
 - 典型的な推薦システムのアルゴリズムにどのような種類のものがあるのか
 - それぞれがどのようにユーザーが好むアイテムを抽出するのか
 - どのような特徴があるのか
 - どのような場面で利用されるのかを直感的に理解することを一番の目的としている
- 種々の具体的なアルゴリズムの実装などの詳細な説明は次章

4.1 推薦アルゴリズムの分類

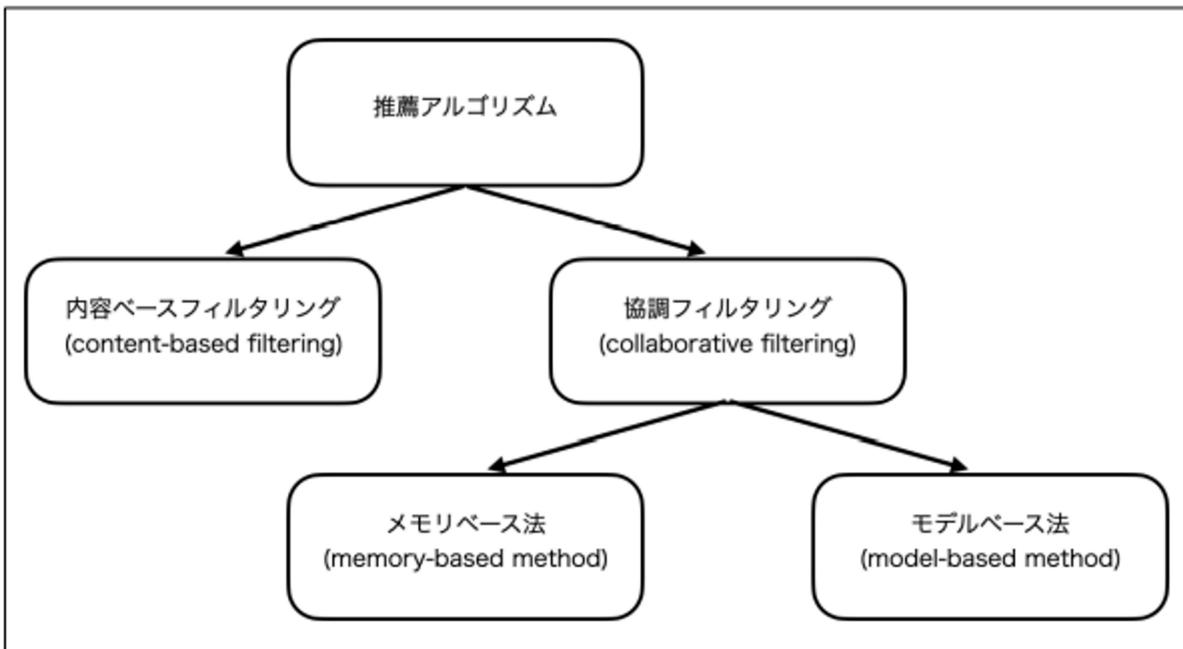


図4-1 推薦アルゴリズムの分類

- 推荐システムのアルゴリズムは図4-1のような分類で説明されることが多い
- 内容ベースフィルタリング (content-based filtering)
 - 本のタイトルや作者、ジャンルなどのようなアイテムの内容を表す情報を利用
 - ユーザーがどのような内容のアイテムを好むかという情報をもとに、内容が似ているアイテムを計算することで推薦を行う
 - たとえば「ユーザー1はミステリーというジャンルが好き」という情報と「本Aのジャンルはミステリーである」という情報をもとに、ユーザー1に本Aを推薦します。
 - このように、推薦を受け取るユーザーの好みに内容が近しいアイテムを探して（計算して）推薦するというのがこのアルゴリズム
 - 推荐対象のアイテムの内容を考慮した上で推薦を行うため「内容ベース」フィルタリングと呼ぶ
- 協調フィルタリング (collaborative filtering)
 - 自分と本の好みが似ている知人に面白かった本を教えてもらうという「口コミ」の過程のように、サービス内の他のユーザーの過去の行動などにより得られる好みの傾向を利用することで推薦を行う
 - たとえば推薦を受け取るユーザーと好みの傾向が似ているユーザーをサービス内で探し出し、そのユーザーが好むアイテムを推薦を受け取るユーザーにも推薦すること
 - サービス内の過去の購買履歴から、ユーザー1とユーザー2は同じような本を好むことが分かっている場合に、ユーザー2がすでに購入しているがユーザー1がまだ購入していない本をユーザー1に推薦するといった形
 - 推荐を受け取るユーザーだけではなく、サービス内の他のユーザーとの協調的な作業によって推薦するアイテムを決定するため、「協調」フィルタリングと呼ぶ
 - 予測の実行方法の観点からメモリベース法 (memory-based

method) と モデルベース法 (model-based method) に分類される

- メモリベース法

- 推薦システムが利用されるまではシステム内のユーザーのデータを蓄積するのみで予測のための計算は行わず、推薦を行うタイミングで蓄積されたデータのうち必要なものすべてを用いて予測計算を行う
- 予測のタイミングで利用するデータをすべてメモリに格納して計算を行うことから「メモリ」ベース法と呼ばれる

- モデルベース法

- 推荐システムが利用される以前にあらかじめシステム内で蓄積されたデータの規則性を学習したモデルを作成しておく、予測時には事前に作成されたモデルと推薦を提供する対象のユーザーのデータのみを利用して計算を行う
- 事前にモデルを作成するので「モデル」ベース法と呼ぶ

4.2 内容ベースフィルタリング

4.2.1 概要

- ユーザーがどのような内容のアイテムを好むかを表す (user profile) と、アイテムのさまざまな性質を表す特徴を抜き出したアイテム特徴 (item feature) との一致度、つまり類似度を計算することで、好みに合ったアイテムをユーザーに推薦するアルゴリズム



図4-2 ユーザープロファイルの例

- ユーザープロファイルはたとえば図4-2のように、それぞれのユーザーが好むアイテムの特徴を並べたリストのような形で表現
 - ユーザー1のユーザープロファイルに注目すると
 - ユーザー1が好きな作者は「鈴木一郎」
 - 好きなジャンルは「ミステリー」
 - 好きな出版社は「A社」



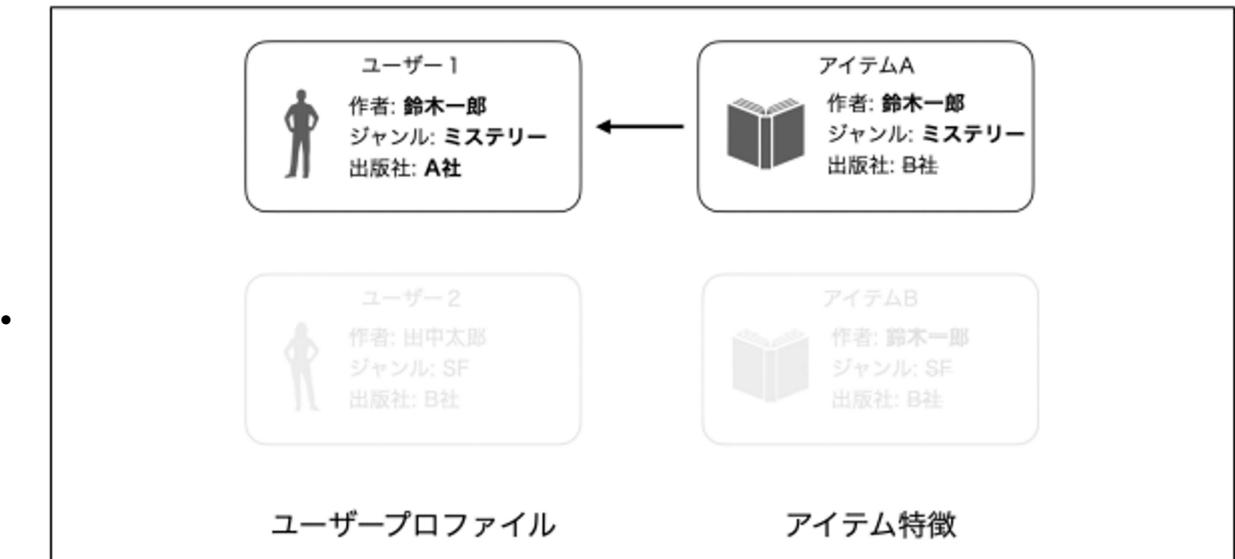


図4-5 より似ているアイテム特徴を持つアイテムをユーザー1に推薦する

- このとき、アイテムAとアイテムBの2つのアイテムのうちアイテムAのアイテム特徴のほうがユーザー1のユーザープロファイルと一致している要素が多く、類似度が高いと考える
 - よって、アイテムAをユーザー1に推薦する（図4-5）。
 - 以上が基本的な内容ベースフィルタリングのアルゴリズムの概要

4.2.2 アイテム特徴の獲得

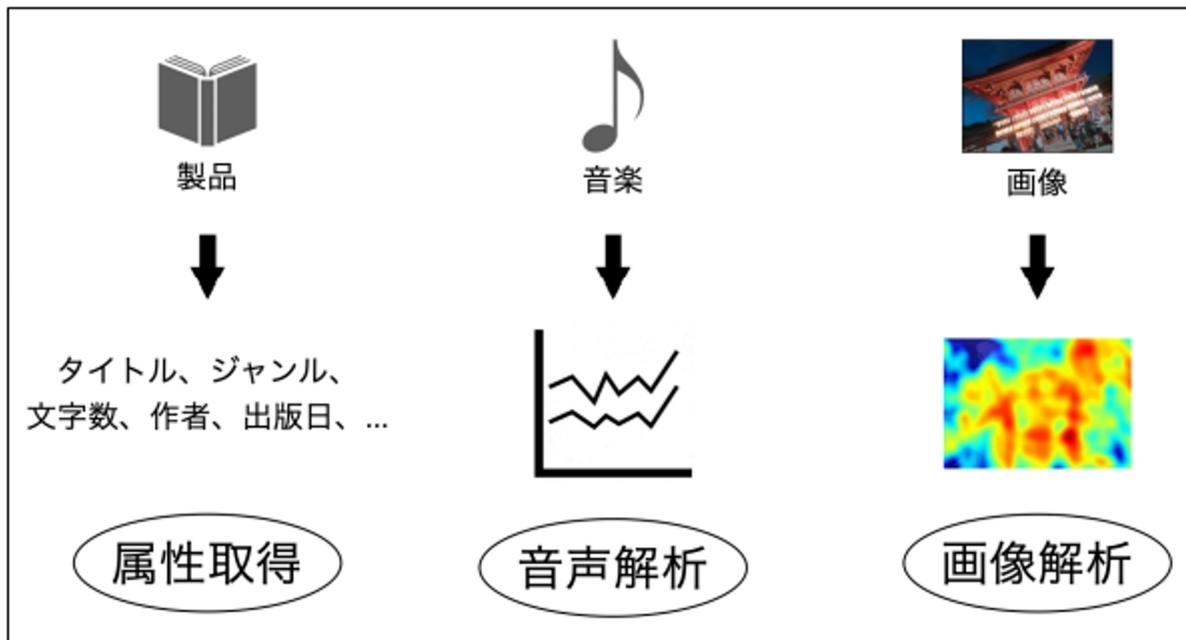


図4-6 アイテム特徴の獲得

- ・ 内容ベースフィルタリングで利用するアイテム特徴の獲得方法
 - ・ 図4-6のように アイテムの性質によってさまざま
 - たとえば本などの製品
 - タイトル
 - ジャンル
 - 文字数

- 作者
- 出版日
- などさまざまなアイテムの属性情報を取得できる
 - アイテムが音楽などの音声データ
 - 作曲者
 - 作曲年
 - などのアイテムの属性情報を取得
 - たとえアイテムが属性情報として持っていない場合でも、音声解析によって音の高さや音色、音量などの情報を取得してアイテム特徴とすることもできる
 - 画像データの場合、たとえば写真ならば写っている物体や撮影された場所、時間などの属性情報に加え、画像解析などの技術を駆使することで色彩情報や被写体の形の特徴などの情報を追加で取得することもできる

4.2.3 ユーザープロファイルの獲得

- ユーザーのアイテムの内容への好みを表すユーザープロファイルの獲得には、大きく2つの方法
 - ユーザーの過去の行動履歴に基づいてユーザープロファイルを作成する間接指定型
 - たとえばユーザーの購入履歴の中で最も多く出現しているアイテムの特徴をそのユーザーが好むものとしてユーザープロファイルを作成（図4-7）
 - この例だと、ユーザーは4つのアイテムを過去に購入している。
 - その中で
 - 作者が「田中太郎」であるものが3つと最も多く
 - ジャンルは「SF」であるものが3つと最も多く
 - 出版社は「B社」のものがこれもまた3つと最も多いため、
 - 「**作者：田中太郎、ジャンル：SF、出版社：B社**」というユーザープロファイルを獲得

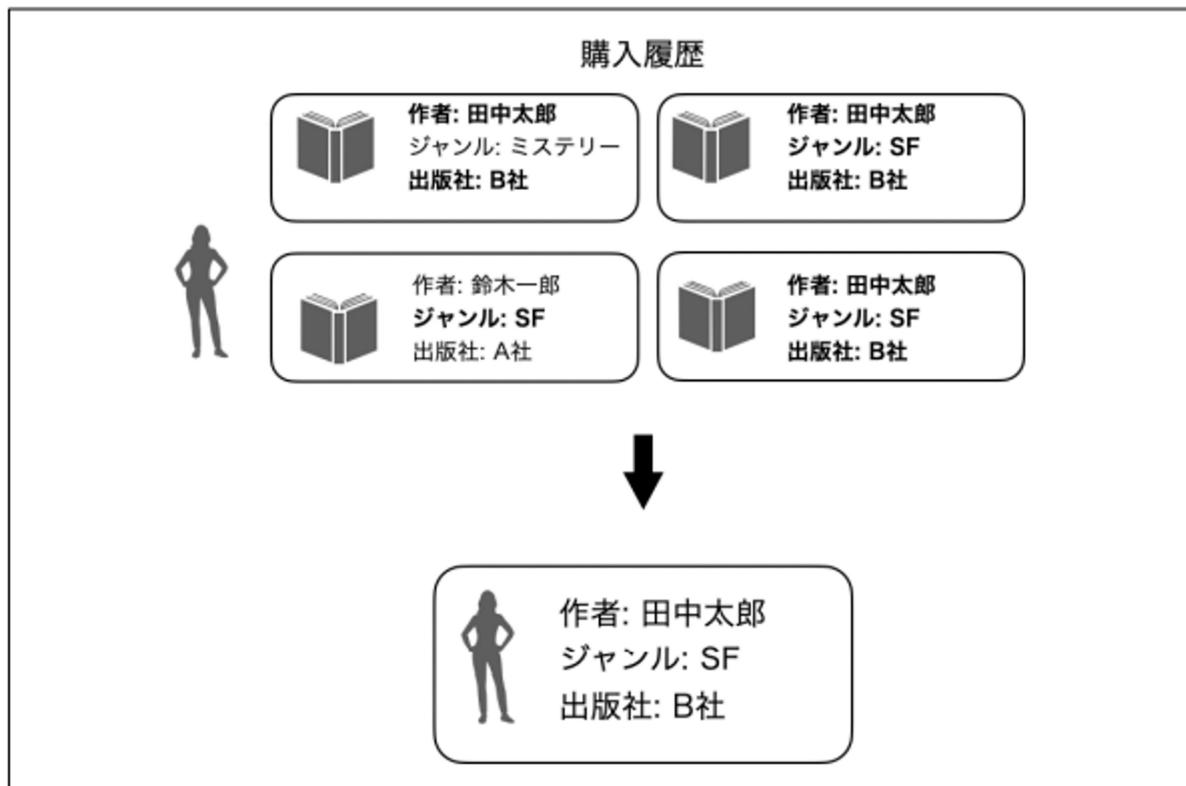
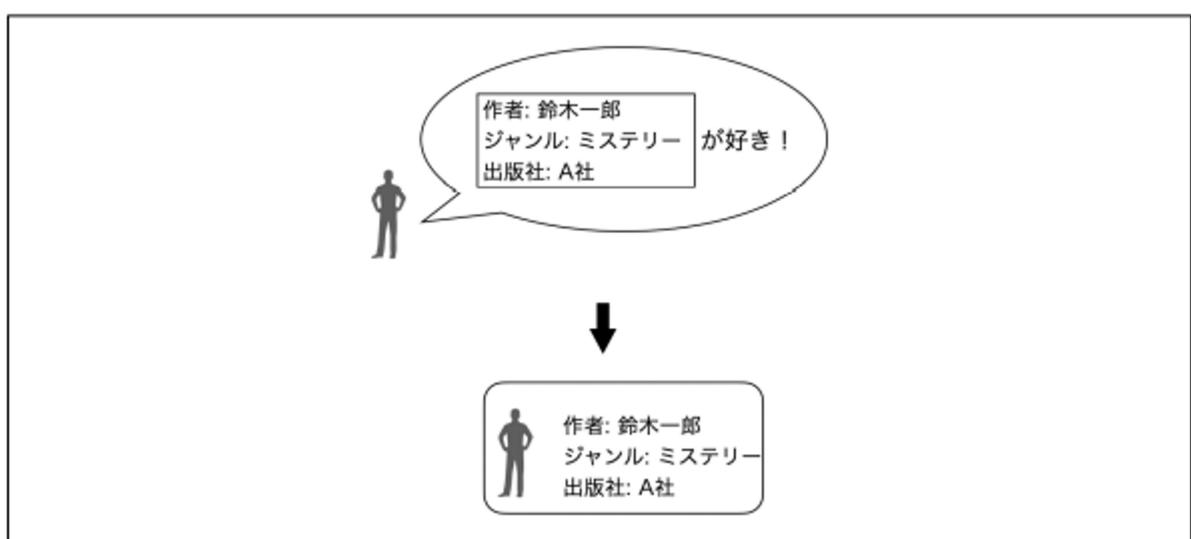


図4-7 ユーザープロファイルの獲得（間接指定型）

- もう1つは、ユーザーに自身が好きなアイテムの特徴を明示的に指定してもらう直接指定型
 - たとえば図4-8では、ユーザーが好きな作者は「鈴木一郎」であり、好きなジャンルは「ミステリー」、好きな出版社は「A社」であると自ら指定しており、**その内容がそのままユーザープロファイルとして獲得**
 - このような明示的な好みの指定は、サービスへのユーザー登録直後のオンボーディングや、マイページなどから行えることが多い



4.3 協調フィルタリング

- 協調フィルタリングにはメモリベース法とモデルベース法の2種類

○ メモリベース法

- 推薦を受け取るユーザーと好みが似ているユーザーに着目して推薦を行う
ユーザー間型メモリベース法 (user-user memory-based method)
- 推薦を受け取るユーザーが好むアイテムと似ているアイテムに着目して推薦を行う
アイテム間型メモリベース法 (item-item memory-based method)

4.3.1 メモリベース法のアルゴリズム概要

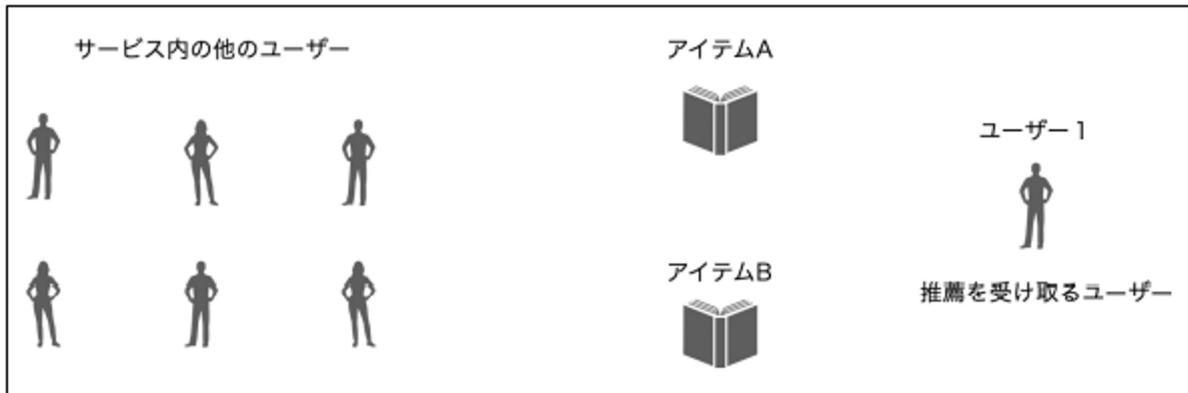


図4-9 ユーザー間型メモリベース法を用いてユーザー1にアイテムを推薦する

- 簡単な例でユーザー間型メモリベース法の協調フィルタリングのアルゴリズムの概要を理解
- 図4-9の右側のユーザー1が推薦を受け取るユーザーであるとする
- 中央にある2つのアイテム、アイテムAとアイテムBのどちらかをこのユーザー1に推薦する

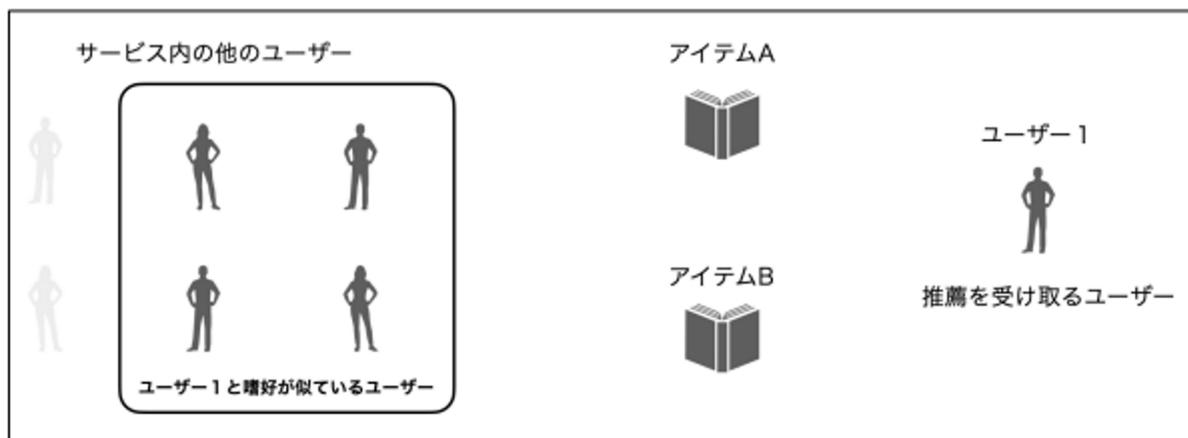


図4-10 ユーザー1と嗜好が似ているユーザーを探し出す

- まずサービス内の他のユーザーの中から、推薦を受け取るユーザーであるユーザー1とアイテムへの好みが似ているユーザーを探し出す
- 今回は、サービス内の他の6人のユーザーのうち4人がユーザー1と好みが似ているユーザーとして探し出された（図4-10）

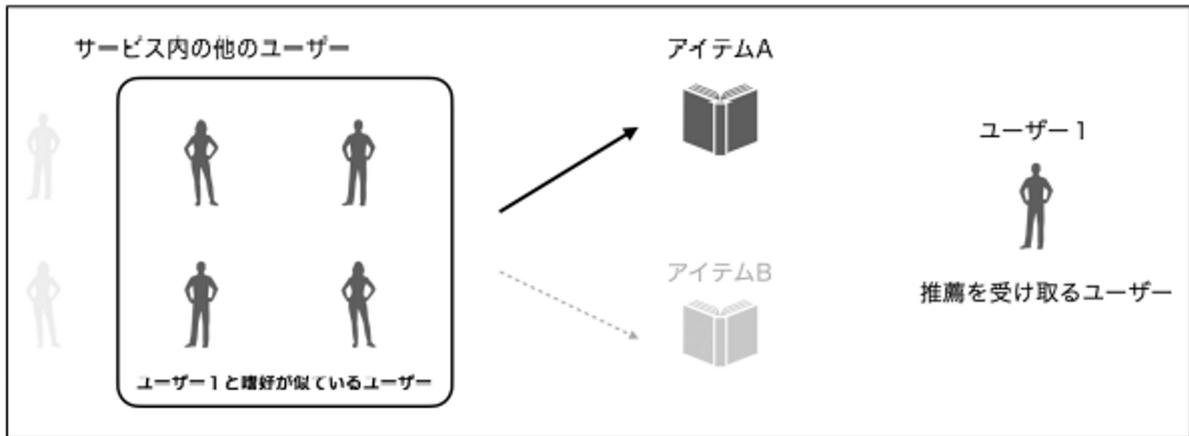


図4-11 ユーザー1と嗜好が似ているユーザーが好むアイテムを探し出す

- 次に、このユーザー1と好みが似ているユーザーたちが好むアイテムを、ユーザー1へと推薦する候補のアイテムの中から探し出す
- ユーザー1と好みが似ているユーザーたちは、アイテムBよりも **アイテムAのほうを好む** ものとする（図4-11）。

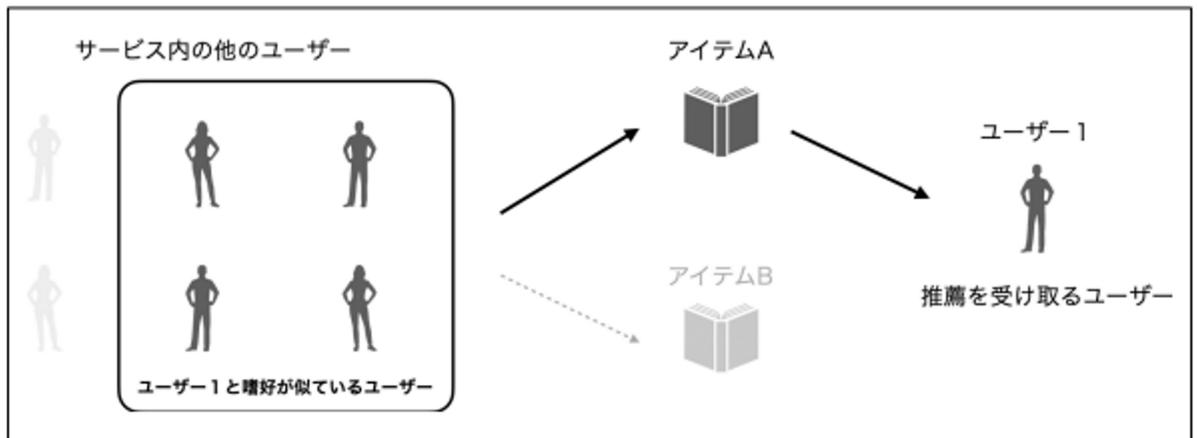


図4-12 ユーザー1と嗜好が似ているユーザーが好むアイテムをユーザー1に推薦する

- 最後に、ユーザー1と好みが似ているユーザーが好むアイテムをユーザー1に推薦する（図4-12）
- 以上が、基本的なユーザー間型メモリベース法のアルゴリズムの概要
- 推薦を受け取るユーザーと好みが似ているユーザーをどのように探し出すのかについてもう少し詳しく説明する
- ここでは、ユーザーの購買履歴からユーザーのアイテムの好みを推測することで、好みの傾向が似ているユーザーを探し出す
 - たとえばユーザーが過去に購入したアイテムは好んでおり、購入しなかったアイテムは好みないと考えることができる。このように、アイテムの具体的な属性情報などを利用しないのは協調フィルタリングの大きな特徴

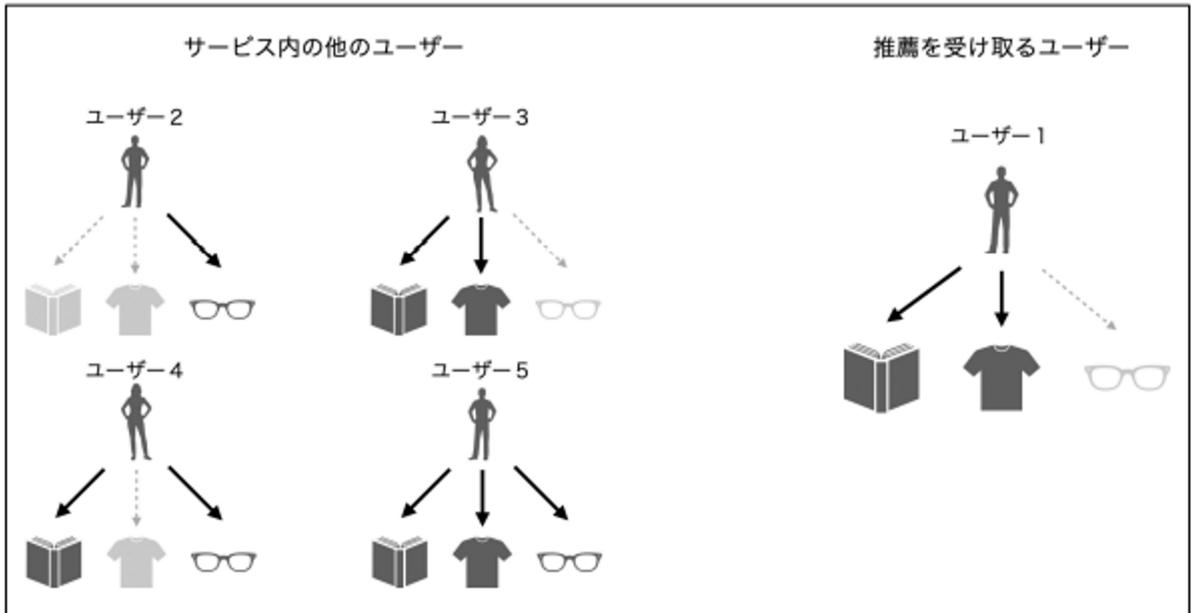


図4-13 過去の購買履歴から嗜好が似ているユーザーを探し出す

- 図4-13の例で考えてみる
- 推薦を受け取るユーザーであるユーザー1は、**本とTシャツ**を購入して眼鏡を購入していない
- このときユーザー1は、**本とTシャツは好み**が眼鏡は好みないということになる
 - このユーザー1と似た好みの傾向を持つユーザーをサービス内の他のユーザー（ユーザー2、3、4、5）の中から探し出す
 - ユーザー2に注目すると、本とTシャツを購入せずに眼鏡だけ購入している。これはユーザー1とは正反対の購買履歴を持っており、好みの傾向は全く似ていない
 - ユーザー3に注目すると、3つすべてのアイテムの購買履歴が一致しており、ユーザー1と全く同じ好みの傾向を持つと言える
 - ユーザー4は本を購入したというのは一致していますが、他の2つのアイテムについては逆の傾向を持つ
 - ユーザー5は本とTシャツの2つのアイテムへの傾向が一致

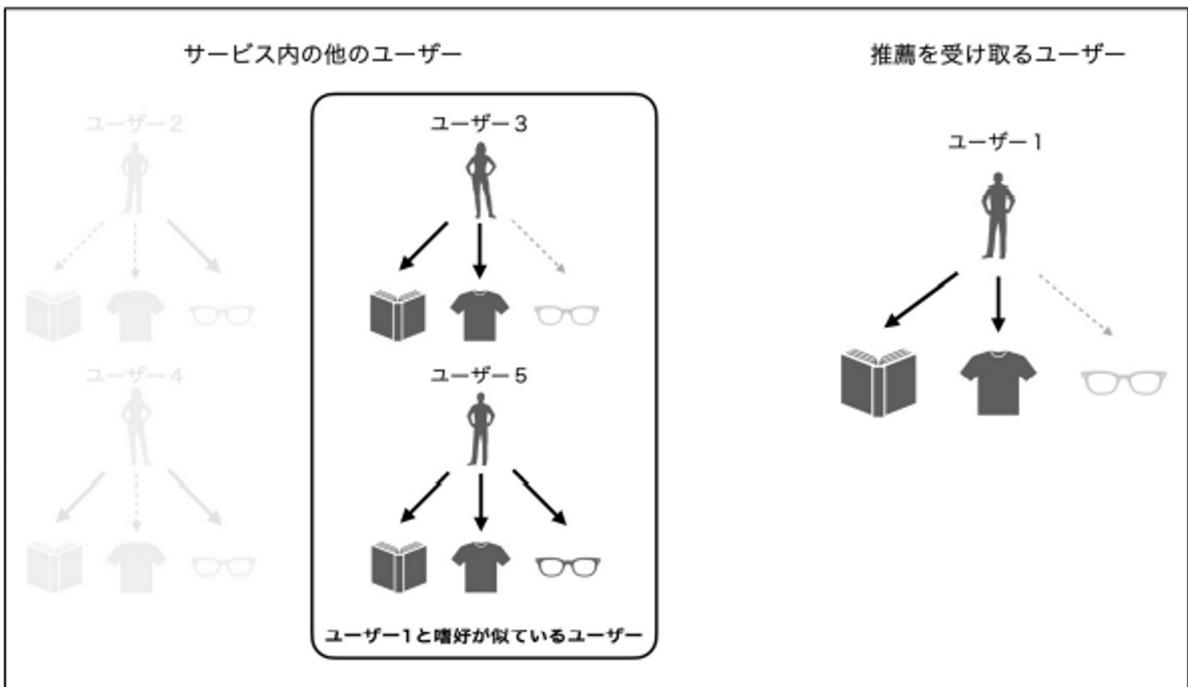


図4-14 過去の購買履歴がユーザー1と似ているユーザーを嗜好が似ているユーザーとする

- ここでは、2つ以上のアイテムの好みの傾向が一致しているユーザーをユーザー1と好みが似ているユーザーであると考える
- このとき **ユーザー3とユーザー5がユーザー1と好みが似ている** ユーザーとして探し出されることとなる（図4-14）

4.3.2 嗜好データの獲得と評価値行列

- 協調フィルタリングでは、サービス内の他のユーザーの過去の行動履歴などから好みの傾向を推定して推薦に利用した
- このようなユーザーからアイテムへの好みの情報を**嗜好データ (preference data)** と言う
- 嗜好データを獲得する方法には大きく2つに分かれる
 - 明示的獲得 (explicit feedback)
 - ユーザーに5段階評価でアイテムのレビューをしてもらったり
 - 好きなアイテムのジャンルなどを聞いて回答してもらったり
 - など、ユーザーにアイテムの好き嫌いや関心のあるなしを質問し回答してもらうことで嗜好データを獲得する方法
 - 暗默的獲得 (implicit feedback)
 - ユーザーがアイテムを購入
 - お気に入りに登録、閲覧
 - といったような、サービス内におけるユーザーの行動履歴からアイテムに対する関心を推定して嗜好データとする方法

| | アイテムA | アイテムB | アイテムC | アイテムD | アイテムE | アイテムF |
|-------|-------|-------|-------|-------|-------|-------|
| ユーザー1 | - | 3 | 4 | 4 | - | - |
| ユーザー2 | - | 2 | - | - | - | 3 |
| ユーザー3 | 3 | - | - | 5 | - | - |
| ユーザー4 | - | - | - | 4 | - | - |
| ユーザー5 | 2 | - | 1 | 2 | - | - |

図4-15 評価値行列の例（評価値は1から5の5段階評価であり、大きいほど良い評価である）

- ・収集した嗜好データなどに基づき、ユーザーがアイテムをどれほど好むのかという嗜好の度合いを定量的に表したものを見たものを評価値（ratings）と言う
- ・その評価値を成分として持つユーザー×アイテムからなる図4-15のような行列を見た評価値行列（rating matrix）と言う
- ・典型的な推薦システムの問題設定では、評価値行列が与えられた上で対象ユーザーのあるアイテムへの未知の評価値を予測計算する
- ・この例では評価値は1から5の5段階評価であり、数字が大きいほど良い評価であるとする
- ・たとえばユーザー1のアイテムCへの評価値は4、アイテムBへの評価値は3。つまり、ユーザー1はアイテムBよりアイテムCのほうを好むということが分かる
- ・評価値を予測するのはユーザーが未評価なアイテムであり、この図では「-」という記号で表している
 - たとえばユーザー1からアイテムEへの評価値は未評価
 - 予測評価値をサービスで利用する際は、ユーザーが閲覧しているアイテムの補足情報として予測評価値をそのまま表示する場合や、予測評価値に基づき並び替えられたアイテムのリストを表示する場合がある
 - アイテムのリストを表示する際、一般的にはユーザーがすでに評価しているアイテム（購入済みのようなアイテム）はリストに含めない

4.3.3 モデルベース法のアルゴリズム概要

- ・協調フィルタリングの中でもモデルベース法とは、既知データの規則性を学習したモデルを事前に作成しておくことで、未知のアイテムへの評価値を予測して推薦する方法
- ・予測に使うモデル
 - クラスタリングを使ったモデルや
 - 回帰問題や分類問題として評価値を直接予測するモデル
 - 線形回帰などの回帰モデルを利用して、過去のユーザーからアイテムへの評価値データを訓練データとしてモデルを学習することで、未知のアイテムへの評価値を予測
 - トピックモデル（Topic Model）を利用するもの
 - LDA（Latent Dirichlet Allocation）などの手法を適用して評価値行列を次

元圧縮することで、「SFが好き」などの潜在的な意味を表現する情報を獲得して推薦に用いる

- 行列分解 (Matrix Factorization) を利用するものなどさまざま
 - 評価値行列をその積が元の行列ができるだけ再現するような形でユーザー行列とアイテム行列に分解。分解された行列から得られるユーザーべクトルとアイテムベクトルの類似度計算によって、任意のアイテムへの予測評価値を計算する
 - あるいは、分解から得られるユーザーべクトルやアイテムベクトルを他のアルゴリズムの入力として使うこともある

4.3.4 メモリベース法とモデルベース法の協調フィルタリングの比較

- 協調フィルタリングの2つの手法であるメモリベース法とモデルベース法について「推薦にかかる時間」と「運用性」という2つの観点で簡単に比較
- 「推薦にかかる時間」という観点で比較
 - メモリベース法
 - 推薦時に毎回すべてのデータを使って似ているユーザーやアイテムを探してから、それらを用いて予測を行うため時間がかかる
 - モデルベース法
 - 事前にデータの規則性を捉えたモデルを作成しているため、作成済みモデルを利用した予測だけ行えばすぐに推薦を行うことができるので時間がかかるない
- 「運用性」の観点で比較
 - メモリベース法
 - ユーザーやアイテムのデータに変更があろうとも、推薦のたびに毎回すべてのデータを利用するので特別な考慮なく常に最新のデータを反映させた推薦を行う
 - モデルベース法
 - ユーザーやアイテムのデータに変更がありその変更を活かした推薦を行うためには、再度モデルを作り直す必要がある
 - モデルの更新タイミングや再学習にかかる計算コストなどを考える必要がある
 - 運用が比較的難しい

4.4 内容ベースフィルタリングと協調フィルタリングの比較

- 内容ベースフィルタリングと協調フィルタリングの2つのアルゴリズムの比較
- それぞれのアルゴリズムの特徴について説明
- ここでは神鳶敏弘の「推薦システムのアルゴリズム」を参考にして、BalabanovicやR. Burkeなどに基づいて以下の表4-1のように7つの観点における比較を行う
- 推荐システムの研究が進んだ現在では、それぞれのアルゴリズムが苦手とされていた観点に対してもさまざまなアプローチが提案されてはいるが、ここでは標準的な協調

フィルタリングと内容ベースフィルタリングの性質について比較する

表4-1 協調フィルタリングと内容ベースフィルタリングの比較

| | 協調フィルタリング | 内容ベースフィルタリング |
|---------------------|-----------|--------------|
| 多様性の向上 | ○ | × |
| ドメイン知識を扱うコスト | ○ | × |
| コールドスタート問題への対応 | × | △ |
| ユーザー数が少ないサービスにおける推薦 | × | ○ |
| 被覆率の向上 | × | ○ |
| アイテム特徴の活用 | × | ○ |
| 予測精度 | ○ | △ |

4.4.1 多様性の向上

- ・ 多様性とは、推薦結果に含まれるアイテムが互いに似ていないことである（「3.4.2 目新しさ・セレンディピティ・多様性」で説明）
- ・ 推薦結果の多様性を高めたい場合、内容ベースフィルタリングよりも協調フィルタリングのほうが有効
- ・ 協調フィルタリング
 - たとえ推薦を受けるユーザー自身が知らないようなことでも、サービス内の他のユーザーが知っており評価していればその情報をもとに推薦を行うことができる
 - たとえば、推薦を受けるユーザーが読んだことのないジャンルの本や知らない新人作者の本であってもそれらを推薦の対象にすることができる
- ・ 内容ベースフィルタリング
 - 推荐を受け取るユーザーのユーザープロファイルはそのユーザーが知らない作者やジャンルの情報を反映することができない
 - 未知のアイテムなのでサービス内の過去の嗜好情報にもない
 - もちろん直接嗜好を指定することもできないため
 - そのため、推薦を受け取るユーザーが知らない情報をアイテム特徴として持つアイテムとユーザープロファイルの類似度が高くなりにくく、推薦することが難しい
- ・ よって、**推薦を受けるユーザーが知らない情報を利用して推薦できるという点**から、内容ベースフィルタリングよりも**協調フィルタリングのほうが推薦結果の多様性を高めやすい**と言える

4.4.2 ドメイン知識を扱うコスト

- ・ サービスの運営者が推薦システムの構築時にかかるコストの観点で2つのアルゴリズムを比較
 - 本のジャンルや出版社、ユーザーの性別や年齢といったサービス内のアイテムやユーザー固有の知識・コンテキストといったドメイン知識を適切に扱ったり管理したり
- ・ 協調フィルタリングのほうが低コスト

・ 協調フィルタリングでは、推薦を受け取るユーザーと過去の嗜好データに基づいて似ていると判断されたユーザーがどのようなアイテムを好むのかという情報をもとに推薦を行うため、アイテム自身の内容についての情報やユーザー自身の属性情報などは基本的に不要でした。一方で内容ベースフィルタリングでは、アイテム特徴やユーザーープロファイルの作成において、適切にドメイン知識を利用しなければ良い推薦が行えません。たとえば内容ベースフィルタリングの説明を行った「4.2 内容ベースフィルタリング」では、アイテム特徴やユーザーープロファイルの作成において作者や出版社といった知識を利用しました。もしここでこれらの知識がなく、ジャンルのみで作成したとしましょう。その場合、ジャンルさえ一致していればそのアイテムを推薦することになり、本当にユーザーの関心があるアイテムを選びにくくなってしまいます。もちろん、適切にドメイン知識を扱って、できるだけユーザーが好むであろうアイテムを推薦できるようにすることが望ましいのですが、前提として、さまざまなドメイン知識を管理するデータベースの維持管理は高コストであることと、そのドメイン知識をうまく扱えなければ良い推薦ができないということ自体も大きなコストになっ 4.4 内容ベースフィルタリングと協調フィルタリングの比較 75 てきます。このように、サービスの運営側としては、運用の観点からなるべくドメイン知識を扱わずに推薦を行いたいといった事情もあります。この点においては、ドメイン知識をうまく扱わなければ良い推薦ができない内容ベースフィルタリングよりも、ドメイン知識がなくともユーザーの行動履歴があれば良い推薦が実現できる協調フィルタリングのほうが望ましい性質を持っていると言えるでしょう。

4.4.3 コールドスタート問題への対応

コールドスタート問題とは、サービス内でのユーザーやアイテムに関する情報が少ないケース、特に新規ユーザーや新規アイテムについて適切に推薦を行うことが難しいという問題のことです。コールドスタート問題への対応という観点では、内容ベースフィルタリングのほうが望ましい性質を持っていると言えるでしょう。ユーザーの過去の嗜好データがなければ推薦ができない協調フィルタリングにおいて、コールドスタート問題は深刻なものとなります。サービスに登録したばかりの新規ユーザーはサービス内に過去の行動履歴などの情報が少ないため、嗜好の似ているユーザーを探すことが難しいからです。新規アイテムについても同様に、サービス内でユーザーにクリックされたり購入されたりすることでユーザーからの嗜好データが集まっていなければ、協調フィルタリングで推薦を行うことができません。一方で内容ベースフィルタリングでは、アイテム特徴やユーザーープロファイルさえ獲得することができれば推薦を行うことができます。新規ユーザーであっても明示的にアイテムへの嗜好を入力してもらうことができれば、最初からユーザーが関心を持ちそうなアイテムを推薦することができます。新規アイテムであってもアイテム特徴が与えられていれば、ユーザーープロファイルに基づいてユーザーに推薦することができます。しかし、サービスを使い始めたばかりのユーザーにいきなり十分な嗜好情報を入力してもらってユーザーープロファイルを獲得することは困難であることから、内容ベースといえどもコールドスタート問題に完全に対応できるわけではありません。以上より、協調フィルタリングよりは内容ベースフィルタリングのほうがコールドスタート問題への対応を行いやすいですが、内容ベースフィルタリングを用いたとしても十分に対応するのは難

しいと言えます。

4.4.4 ユーザー数が少ないサービスにおける推薦

システムを利用するユーザー数が少ない場合でも適切な推薦を行うことができるか、という観点で2つのアルゴリズムを比較してみます。新規サービスなどにおいて 76 4章 推薦アルゴリズムの概要は特に重要な観点となるでしょう。この点では内容ベースフィルタリングのほうが望ましい性質を持っていると言えます。サービス内の他のユーザーの行動履歴をもとに推薦を行う協調フィルタリングにおいては、ユーザー数が少ないと十分に似ているユーザーを獲得したり、似ているユーザーの行動履歴から推薦対象のユーザーが関心を持つであろうアイテムを特定することが難しいために、良い推薦を行うことが難しいです。さらに、良い推薦を行うことができないとユーザーはサービスに価値を感じてくれなくなり、ユーザー数が増えなくなってしまいます。そして、ユーザー数が増えないといつまでたっても良い推薦ができないといった悪い流れがでてしまします。一方で内容ベースフィルタリングでは、アイテム特徴やユーザープロファイルさえ獲得することができれば推薦を行うことができるので、システム内にユーザーがどれくらいいるかということはあまり推薦に関係ありません。よって、まだ利用ユーザー数が少ないサービスにおいては、協調フィルタリングよりも内容ベースフィルタリングを使ったほうが良い推薦を行いやすいと言えるでしょう。

4.4.5 被覆率の向上

被覆率 (coverage) とは、サービスにあるすべてのアイテムのうち、推薦システムでユーザーに推薦することができるアイテムの割合のことです。被覆率が少ない状態とは一部のアイテムに偏った推薦がなされている状態であり、サービスの利用ユーザーにとっても提供者にとっても良い状態とは言えないです。この観点では、内容ベースフィルタリングのほうが望ましい性質を持っていると言えるでしょう。システム内の他のユーザーの行動履歴をもとに推薦を行う協調フィルタリングでは、推薦を受け取るユーザーと似ているユーザーが誰も試していない、つまり評価していないアイテムについては推薦することができません。この点において、サービスにあるアイテムのうち推薦できるアイテムは限られたものとなってしまいます。一方で内容ベースフィルタリングでは、システム内のユーザーのユーザープロファイルと関連するアイテム特徴を有するアイテムであればサービスに存在するどのアイテムでも推薦することができます。適切にアイテム特徴を管理してユーザープロファイルに反映されるようにすることで、全く推薦されないアイテムをできるだけなくすことが可能でしょう。よって、被覆率を向上させるという点では、協調フィルタリングよりも内容ベースフィルタリングのほうが望ましい性質を持っていると言えるでしょう。4.4 内容ベースフィルタリングと協調フィルタリングの比較

4.4.6 アイテム特徴の活用

次に、アイテム特徴をうまく活用した推薦が行えるかという観点で比較します。この観点では、内容ベースフィルタリングが望ましい性質を持っていると言えます。ドメイン知識を利用せずにユーザーの過去の嗜好データのみに基づいて推薦を行う協調フィルタリングにおいては、服の色が何色であるかなどのアイテムの属性情報を基本的には考慮すること

ができません。同じ商品のサイズや色の違うもの、あるいは同じ目的の競合商品などもまったく異なるアイテムとして扱われます。そのため、たとえばある黒い服が推薦されたがその服が（その色にかかわらず）気に入らなかったとします。しかし、自分と嗜好が似ているユーザーが色違いの同じ服を好んでいる場合、次にそれらが自分に推薦されてしまうというようなことが起こります。一方で内容ベースフィルタリングにおいては、アイテムのさまざまな特徴を明示的に考慮した上で推薦を行うので、色違いの同じ服は同時に推奨しないということや、ユーザー好みの色が分かった際に同じ色の異なる種類の服を推薦するなどといったことが柔軟に行えます。よって、アイテム特徴の活用という観点では、協調フィルタリングよりも内容ベースフィルタリングのほうがうまく扱って推薦できると言えるでしょう。

4.4.7 予測精度

最後に、ユーザーの評価値への予測精度についても触れておきます。ここまでさまざまな観点で協調フィルタリングと内容ベースフィルタリングを比較してきました。これまでの説明の通りそれぞれのアルゴリズムには得意不得意があり、常にどんな状況でも一方のアルゴリズムのほうが予測精度の観点で望ましいということは言えません。ここでは、ある程度の規模のサービスにおいて、多数派である一定以上アクティブにサービスを利用している一般的な嗜好傾向を持つユーザーに対する推薦を考えます。この場合、内容ベースフィルタリングよりも協調フィルタリングのほうが高い精度で予測を行うことができる一般的に言われています。獲得されたユーザープロファイルやアイテム特徴にそのまま基づいて推薦を行う内容ベースフィルタリングよりも、さまざまなユーザーの行動履歴を推薦結果に反映できる協調フィルタリングのほうが複雑なユーザーの嗜好を考慮できるからであると考えられるからです。別の観点として、ユーザーがサービスを利用し続けるほど推薦に活かせるデータは増え続け、さらなる予測精度の向上が見込まれます。この点でも、自身の嗜好情報であるユーザー78章 推薦アルゴリズムの概要 ザーブロファイルだけでなく、サービス内の他のユーザーの嗜好データも利用して推薦を行える協調フィルタリングのほうが、サービス内のデータが増えることの恩恵を受けやすいと考えられます。

4.4.8 比較のまとめ

7つの観点から協調フィルタリングと内容ベースフィルタリングの比較を行なながら、それぞれのアルゴリズム特徴について説明しました。多様性の向上とドメイン知識を扱うコスト、一般的な予測精度の観点では協調フィルタリングのほうが、コールドスタート問題への対応、ユーザー数が少ないサービスにおける推薦、被覆率の向上、アイテム特徴の活用の観点では内容ベースフィルタリングのほうが望ましい性質を持ちます。このように、それぞれのアルゴリズムには得手・不得手がありますので、解決したいタスクの性質に応じて適切なアルゴリズムを選択する必要があります。ただ、アルゴリズム選択の際にはどちらか一方だけを選択しなければならないということではなく、状況によって使い分けたり、そもそもこの2つをいい具合に組み合わせることでいいところを取りをしようとする手法も存在します。もちろん、アルゴリズムが複雑になればなるほど、実装や運用のコストは高くなってしまうので、その辺りのバランスを鑑みながら、さまざまな選択肢の中から適切なアルゴリズムを選択できるようになることが重要でしょう。

4.5 推薦アルゴリズムの選択

ここまで推薦アルゴリズムの分類とそれぞれの特徴について説明してきました。次に、適切な推薦アルゴリズムを選択するための指針を紹介します。内容ベースフィルタリングか協調フィルタリングかの選択指針について説明します。前節では、2つの手法を比較しながらそれぞれの長所と短所について説明しました。推薦システムを導入するサービスで達成したい目的に応じて、これらの特徴の違いを加味した上で適切なものを選択していくこととなります。前提として、これら2つのアルゴリズムからどちらか一方のみを選択しなければならないわけではなく、適切に組み合わせたハイブリッドな手法を取ることが多いです。また、1つのシステムに対して1つのアルゴリズムだけを適用するというわけでもなく、ユーザーやアイテムの状況に応じて複数のアルゴリズムを使い分けたり変化させたりする必要があります

まず一般的な指針として、サービス内のデータ量に応じてアルゴリズムを選択することが考えられます。前節の「システムのユーザー数」の項目で説明しましたが、新規サービスではユーザー数が少ないために協調フィルタリングを適用するのが困難でした。また、「コールドスタート問題」の項目で説明した通り、新規ユーザーや新規アイテムについてはサービス内でのデータが少ないので、協調フィルタリングの適用が困難でした。そのため、データが少ないので新規サービスにおける推薦や、ある程度の規模のサービスでも新規ユーザーへの推薦や新規アイテムを推薦したい際には、内容ベースフィルタリングを選択することが多いです。一方で、ある程度ユーザー数が増えたりユーザーやアイテムごとのデータが蓄積されたあとは、より複雑なユーザーの嗜好を表現可能な協調フィルタリングのほうが高い精度でアイテムを推薦できることが多いです。そのため、一定のデータが使えるようになった時点で内容ベースフィルタリングから協調フィルタリングにアルゴリズムを切り替えるといった方策が取られることが多いように思います。また、3章でも紹介した推薦システムの提供形態に応じたアルゴリズムの選択も考えられます。概要推薦であれば、サービス全体のユーザーの行動履歴に基づく統計データを用いた人気リストであったり、アイテムの属性などでフィルタリングをかけるようなシンプルな直接指定型の内容ベースフィルタリングが良いでしょう。関連アイテム推薦では、予測の精度の高さやアイテム特徴を管理しなくても良いというコスト面の観点からアイテム間型メモリベース法の協調フィルタリングを選択したり、単純にアイテムの特徴を用いて似ているアイテムを見つけて推薦を行うことが多いです。通知サービスやパーソナライズでは、それぞれのユーザーについての情報をある程度蓄積していることが前提となることが多いため、ユーザー間型メモリベース法の協調フィルタリングや間接指定型の内容ベースフィルタリングをまずは選択することが多いように思います。また、十分なデータが蓄積されている場合は、モデルベース法の協調フィルタリングが有効でしょう。

多様性の観点に応じてアルゴリズムを選択することも重要です。たとえばECサイトにおいて、特定の色で特定のペン先の太さのボールペンを購入したくて探している場合に、違う色や違う太さのボールペン、あるいは鉛筆やシャープペンシルを推薦することはあまり必要とされないでしょう。このような実用品などの、欲しいものがかっちりと決まっているようなアイテムに関しては、あまり多様性の重要度が高くないために堅実にユーザーが欲

しいものを推薦できる直接指定型の内容ベースフィルタリングを利用したほうが良いでしょう。一方で、何かお菓子が欲しくてECサイトを利用しているユーザーについては、これまでの購入履歴などから推測した新しいお菓子を推薦してくれる。このような嗜好品などの、欲しいものが明確に決まっているわけではないアイテムに関しては、多様性が重要となるケースも多いです。そのため、新規性やセレンディピティを重視して多様性を実現できる協調フィルタリングを選択することも多いです。

4.6 嗜好データの特徴

適切な推薦システムを設計するためには、推薦システムの入力となるユーザーの嗜好データを適切な形でサービスで得られるようにすることが必要です。そのためには、嗜好データそのものの特徴について理解しておくことは不可欠なので、本章の最後に付け加えておきます。

4.6.1 明示的獲得と暗黙的獲得の比較

嗜好データの獲得方法には明示的獲得と暗黙的獲得の2つがありました。それぞれの方法で得られた嗜好データには異なる特徴があるため、推薦システムの目的に応じて使い分ける必要があります。ここでは、2つの嗜好データの獲得方法を4つの観点で比較します。それぞれの長所と短所を表4-2にまとめました。

表4-2 明示的獲得と暗黙的獲得の比較

| 明示的獲得 | | 暗黙的獲得 |
|------------|-------|-------|
| データ量 | ×：少ない | ○：多い |
| データの正確さ | ○：正確 | ×：不正確 |
| 未評価と不支持の区別 | ○：明確 | ×：不明確 |
| ユーザーの認知 | ○：認知 | ×：不認知 |