

# 協調フィルタリングによる プロジェクトデータ分析

EPM データを用いた協調フィルタリングの応用例

大杉 直樹, 松本 健一

奈良先端科学技術大学院大学  
情報科学研究科

- **協調フィルタリングの手順**
  - 手順 1: 類似度計算
  - 手順 2: 予測値計算
  - 手順 3: 推薦作成
- **EPM データを用いた協調フィルタリングの応用例**
  - 類似度計算結果の応用例 (現状把握)
  - 予測値計算結果の応用例 (見積もり)
  - 推薦作成結果の応用例 (情報推薦)
- **まとめと今後の課題**
- **お願い**

# 手順 1: 類似度計算

- 推薦対象ユーザと他ユーザの間の類似度を計算する。

- Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J.,  
“GroupLens: An Open Architecture for Collaborative Filtering of Netnews,”  
*Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work (CSCW'94)*, pp.175-186, 1994.

	書籍 1	書籍 2	書籍 3	書籍 4	書籍 5	
推薦対象ユーザ	5 (好き)	5 (好き)	1 (嫌い)	3 (普通)	? (予測対象)	
ユーザ A	5 (好き)	5 (好き)	1 (嫌い)	? (未評価)	5 (好き)	類似度: 1.0
ユーザ B	? (未評価)	5 (好き)	1 (嫌い)	3 (普通)	5 (好き)	類似度: 0.98
ユーザ C	1 (嫌い)	1 (嫌い)	? (未評価)	5 (好き)	1 (嫌い)	類似度: -0.97

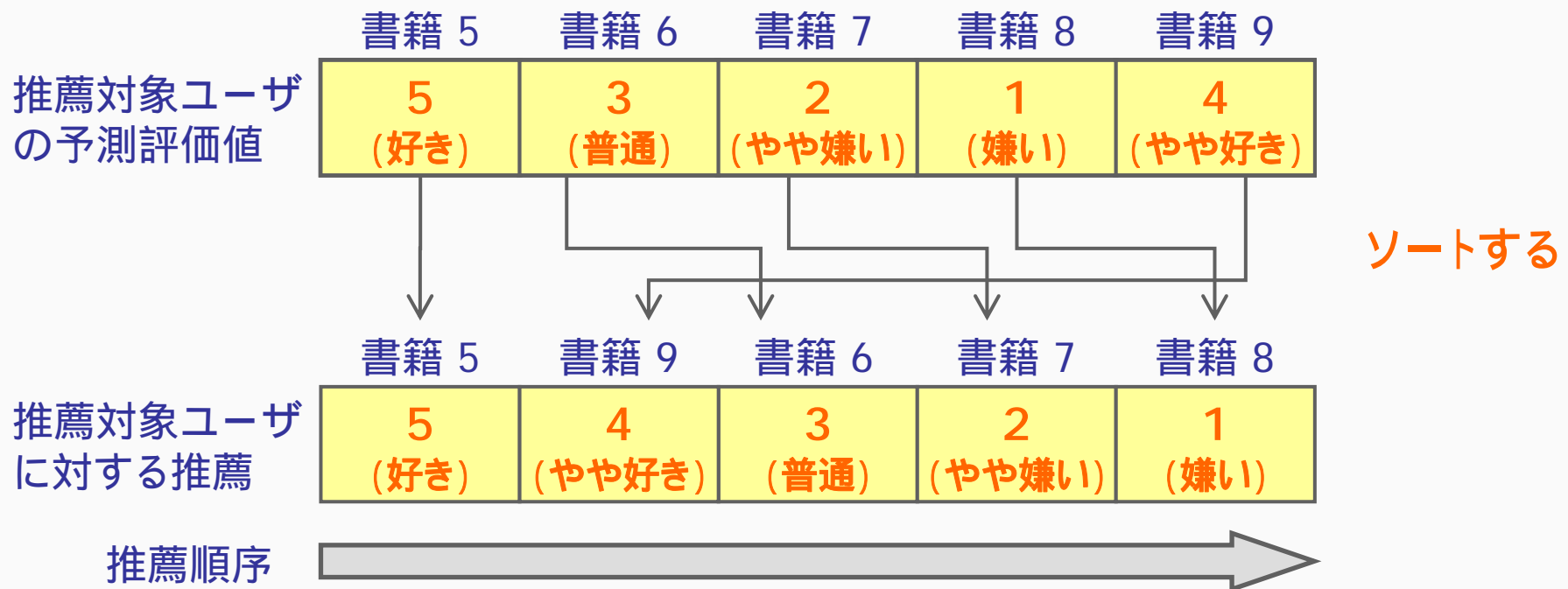
# 手順 2: 予測値計算

- 類似度の高い  $k$  (例えば  $k = 2$ ) 人のユーザの評価を加重平均し, 推薦対象のユーザの評価を予測する.

	書籍 1	書籍 2	書籍 3	書籍 4	書籍 5	
推薦対象ユーザ	5 (好き)	5 (好き)	1 (嫌い)	3 (普通)	5 (好き)	予測する
ユーザ A	5 (好き)	5 (好き)	1 (嫌い)	? (未評価)	5 (好き)	類似度: 1.0
ユーザ B	? (未評価)	5 (好き)	1 (嫌い)	3 (普通)	5 (好き)	類似度: 0.98
ユーザ C	1 (嫌い)	1 (嫌い)	? (未評価)	5 (好き)	1 (嫌い)	類似度: -0.97

# 手順 3: 推薦作成

- 推薦対象ユーザがまだ評価していない書籍全てについて、予測評価値を計算する。
- 予測評価値に関して書籍を降順にソートする。
- 評価値の高い書籍から順にユーザに推薦する。



# EPM データを用いた協調フィルタリングの 応用例

- 協調フィルタリングの過程で得た処理結果を、異なる用途に応用できる。

手順	処理結果の用途	EPM データを用いた 応用例
 <p>手順1. 類似度計算</p>	現状把握	類似性可視化, エキスパート同定 etc.
 <p>手順2. 予測値計算</p>	見積もり	バグ数予測, 規模予測 etc.
 <p>手順3. 推薦作成</p>	情報推薦	リマインダ etc.

- 類似性可視化

- データ中の各要素間の類似性を可視化する。
- デモンストレーション：EPM で収集された「EASE プロジェクト関係者用メーリングリストの投稿履歴」を使って...
  - [EASE プロジェクト関係者間の類似性を可視化](#)する。
  - [メールのサブジェクトに含まれる単語間の類似性を可視化](#)する。

- **エキスパート同定**

- データ中の特定要素に関して知識を持っている人物(エキスパート)を特定する。
- デモンストレーション: EPM で収集された「EASE プロジェクト関係者用メーリングリストの投稿履歴」を使って...
  - メールのサブジェクトに含まれる特定の単語に詳しい人物を特定する。



# 予測値計算結果の応用例：見積もり

- バグ数予測，規模予測 etc.

- データ中の未測定要素を，類似する過去のプロジェクトの値を加重平均して予測する．
- EPM で収集可能な時系列データへの応用も考えられる．

	LOC	サイクロマチック数	開発者数	ファイル総数	総バグ数	
現行 PJ	50k	1000	3	40	400	予測する
過去 PJ A	45k	1000	2	36	360	類似 PJ
過去 PJ B	50k	1100	3	44	440	類似 PJ
過去 PJ C	10k	500	6	20	300	非類似 PJ

- リマインダ

- 開発者が過去に扱った情報の傾向を分析し、開発者が扱うべきであるにも関わらず、見落としていると思われる情報を提示する。
- デモンストレーション：EPM で収集された「EASE プロジェクト関係者用メーリングリストの投稿履歴」を使って...
  - リプライを返すべきであるにも関わらず、見落とされているメールのサブジェクトを提示する。

- 協調フィルタリングの手順について概説した。
- EPM データを用いて協調フィルタリングを行い、その過程で得られる処理結果の応用例を紹介した。

- **応用方法(アプリケーション)の考案**
  - ソフトウェア開発に役立つ協調フィルタリングのアプリケーションを考える。
- **アプリケーションの有用性検証**
  - 実際のソフトウェア開発プロジェクトから収集されたデータを用いて精度評価実験を行い,アプリケーションの有用性を検証する。
- **アプリケーションの実装**
  - 検証の結果,有効だと判断されたアプリケーションを
    - EPM の機能として実装する。
    - 各企業様の環境(データ収集手順や方法)に適合するツールとして実装する。

# お願い1: アンケートにご協力ください

- **お名前と電子メールアドレスのご記入について.**
  - 本名や会社の電子メールアドレスのご記入が難しい方は、ニックネームやフリーメールアドレスをご記入ください。
- **回答方法について**
  - アンケート用紙に記した各キーワードについて、「知っている」或いは「知らない」をお答えください。「知っている」とお答えの場合、どの程度ご興味をお持ちか、4段階でお答えください。
- **アンケート結果の分析について**
  - アンケート結果に協調フィルタリングを適用し、参加者間、並びに、キーワード間の類似関係を可視化します。
  - ご興味をお持ちになると思われるキーワード(お勧めキーワード)を推薦します。推薦結果は、ご記入いただいた電子メールアドレスに個別に連絡いたします。

# お願い2: ご意見をお聞かせください

- ソフトウェア開発に役立つ協調フィルタリングの応用方法として, どんなものが考えられるでしょうか.
- Amazon社の書籍推薦システムで用いられる表データの代わりに, どんなデータが考えられるでしょうか.

	列ラベル1	列ラベル2	列ラベル3	列ラベル4	列ラベル5
行ラベルA	要素 A-1	要素 A-2	要素 A-3	要素 A-4	? (予測対象)
行ラベルB	要素 B-1	要素 B-2	要素 B-3	要素 B-4	要素 B-5
行ラベルC	要素 C-1	要素 C-2	要素 C-3	要素 C-4	要素 C-5
行ラベルD	要素 D-1	要素 D-2	要素 D-3	要素 D-4	要素 D-5

- 今回の分析は, オープンソースプロジェクト NCFE (Naist Collaborative Filtering Engines) で開発されたソフトウェアを用いて行いました.
  - <http://sourceforge.jp/projects/ncfe/>





# 複数プロジェクトから収集されたデータへの適用

- 協調フィルタリングを用いた工数見積もり技法では,
  - 下図のような表データを使用し,
  - プロジェクト(Pj)間の類似度を算出し,
  - 類似する過去 Pj の工数から現行 Pj の工数を予測する.

	設計工数	製造工数	基本設計 欠陥数	詳細設計 欠陥数	試験工数	
現行 Pj	50	20	3	10	40	予測する
過去 Pj A	45	18	2	? (欠損値)	36	類似 Pj
過去 Pj B	? (欠損値)	22	3	11	44	類似 Pj
過去 Pj C	10	10	? (欠損値)	5	30	非類似 Pj

# ミクロ的データへの適用

- 協調フィルタリングを用いた工数見積もり技法では,
  - 下図のような表データを使用し,
  - プロジェクト(Pj)間の類似度を算出し,
  - 類似する過去 Pj の工数から現行 Pj の工数を予測する.

	設計工数	製造工数	基本設計 欠陥数	詳細設計 欠陥数	試験工数	
現行 Pj	50	20	3	10	40	予測する
過去 Pj A	45	18	2	? (欠損値)	36	類似 Pj
過去 Pj B	? (欠損値)	22	3	11	44	類似 Pj
過去 Pj C	10	10	? (欠損値)	5	30	非類似 Pj

# アプリケーションの実装イメージ

# 開発者の視点に立ったアプリケーション



- デモ

- EPM 開発プロジェクトのデータからのナレッジマイニング

# デモ: EPM 開発プロジェクトのデータからの ナレッジマイニング



# お願い1: データをください!

- **必ず実りある分析結果をお返しします。**
  - 協調フィルタリングを適用し, ナレッジマイニングを行います。
- **ご業務に役立つアプリケーションを提案します。**
  - 共同研究契約を結んでいただければ, 提案したアプリケーションをツールとして実装いたします。
- **どんなデータでも構いません。**
  - ご業務(ソフトウェア開発プロジェクトに限られません)の過程で収集されたデータをご提供ください。
- **データを加工する必要はありません。**
  - 欠損値や異常値が沢山含まれていても構いません。
- **秘密は厳守します。**
  - データに触れる者は全員, 守秘義務契約を結びます。

- 類似性可視化

- データ中の各要素間の類似性を可視化する。
- デモンストレーション：EPM で収集された「EASE プロジェクト関係者用メーリングリストの投稿履歴」を使って...
  - EASE プロジェクト関係者間の類似性を可視化する。
  - メールのサブジェクトに含まれる単語間の類似性を可視化する。

	開発者 A	開発者 B	開発者 C	開発者 D	開発者 E
単語 1	Aが1を使用した頻度	Bが1を使用した頻度	Cが1を使用した頻度	Dが1を使用した頻度	? (予測対象)
単語 2	Aが2を使用した頻度	Bが2を使用した頻度	Cが2を使用した頻度	Dが2を使用した頻度	Eが2を使用した頻度
単語 3	Aが3を使用した頻度	Bが3を使用した頻度	Cが3を使用した頻度	Dが3を使用した頻度	Eが3を使用した頻度
単語 4	Aが4を使用した頻度	Bが4を使用した頻度	Cが4を使用した頻度	Dが4を使用した頻度	Eが4を使用した頻度



## • エキスパート同定

- データ中の特定要素に関して知識を持っている人物(エキスパート)を特定する.
- デモンストレーション: EPM で収集された「EASE プロジェクト関係者用メーリングリストの投稿履歴」を使って...
  - メールのサブジェクトに含まれる特定の単語に詳しい人物を特定する.

	単語 1	単語 2	単語 3	単語 4	単語 5
開発者 A	Aが1を使用した頻度	Aが2を使用した頻度	Aが3を使用した頻度	Aが4を使用した頻度	? (予測対象)
開発者 B	Bが1を使用した頻度	Bが2を使用した頻度	Bが3を使用した頻度	Bが4を使用した頻度	Bが5を使用した頻度
開発者 C	Cが1を使用した頻度	Cが2を使用した頻度	Cが3を使用した頻度	Cが4を使用した頻度	Cが5を使用した頻度
開発者 D	Dが1を使用した頻度	Dが2を使用した頻度	Dが3を使用した頻度	Dが4を使用した頻度	Dが5を使用した頻度