協調フィルタリングによる工数，規模，コスト超過の見積りについて

2005年度 第 4 回エンピリカルソフトウェア工学研究会

大杉 直樹

奈良先端科学技術大学院大学 情報科学研究科

EASE プロジェクト研究員

研究の背景

- 定量的データ分析に基づく見積もりは重要！だが,...
- 開発者やプロジェクト管理者の時間も重要.
  - 見積もりを実施するために必要なコストが，得られる利益に見合っていなければならない.

目的とアプローチ

- 目的
  - 下記の特長を備えた簡単に使える(既収集データを最大限に生かせる)定量的見積もり手法／ツールを開発する.
    - 入力データに対する受容性.
    - データ欠損に対するロバスト性(頑健性).
    - 入力データの特性に対する順応性.
- アプローチ
  - データ自動調整機能の実装.
  - 協調フィルタリングの採用.
  - 予測アルゴリズムを交換可能にする設計.

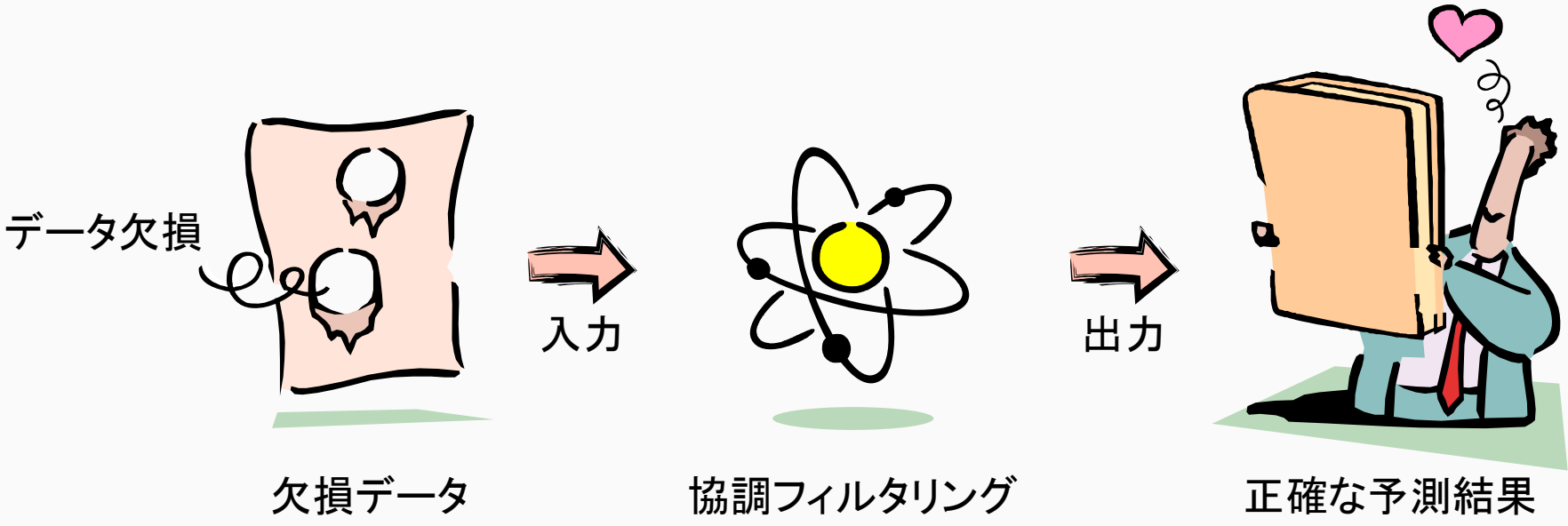
データ自動調整機能の実装

- 見積もりを実行可能な形式にデータを自動調整するようにツールを実装する.
  - 実装したツールは，下記のような CSV（Comma Separated Values）形式のデータを，一切の前処理なしで読み込める.

	Language	Designing cost	Coding cost	# of bugs	...
PJ_A	COBOL	50	20	10	...
PJ_B	COBOL	45	18		...
PJ_C		55	22	11	...
PJ_D	Java	10		30	...
...	...	...	...	...	...

協調フィルタリングの採用

- Amazon 社の書籍推薦システムなどで用いられている予測技法．ユーザの好みの傾向を予測する．
  - 各ユーザが書籍を 5（好き）～1（嫌い）の 5 段階で評価したデータを基に予測する．
  - ユーザが書籍を評価していない部分（データ欠損）が沢山あっても，ユーザの好みを正確に予測できる．



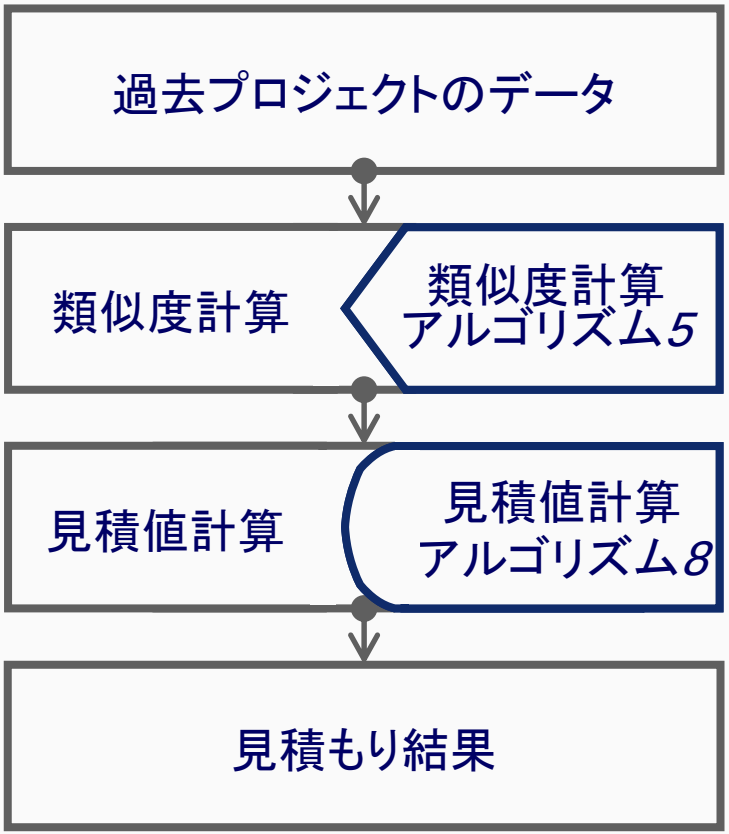
協調フィルタリングに基づく見積もり手順

- ステップ 1：類似度計算
  - 現行プロジェクト（見積もり対象）と過去の各プロジェクトの間の類似度を計算する.
  - 類似度の高い k 個の過去プロジェクトを選ぶ（例えば $k = 2$）.
- ステップ 2：見積値計算
  - 類似プロジェクトの実績値から，現行プロジェクトの見積値を計算する.

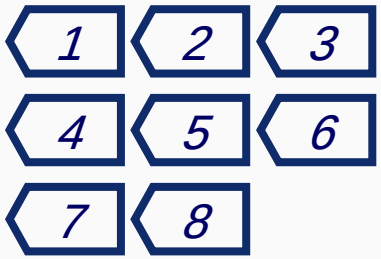
	開発言語	開発種別	概算FP	要員数	開発工数
現行プロジェクト X	Java	新規	40	10	見積結果 37.5
類似度：+1.0 ェクト A	Java	新規	データ欠損	8	40
類似度：+0.9 ェクト B	Java	データ欠損	25	6	35
類似度：−1.0 ェクト C	データ欠損	保守	100	40	250

交換可能な予測アルゴリズム

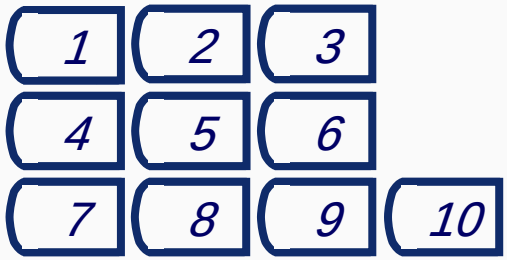
- 見積もりの各ステップは，いくつかのアルゴリズムの中から選んで実行する．データの特徴に応じて，使用するアルゴリズムを切り替える．



類似度計算アルゴリズム：8 種実装済



見積値計算アルゴリズム：10 種実装済



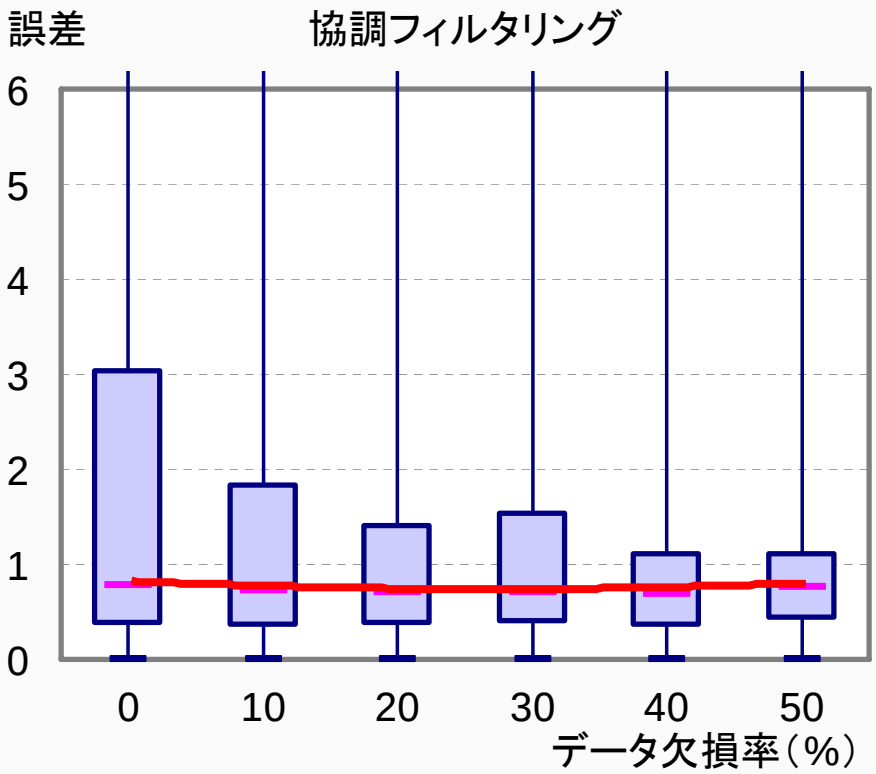
これまでに発表した文献

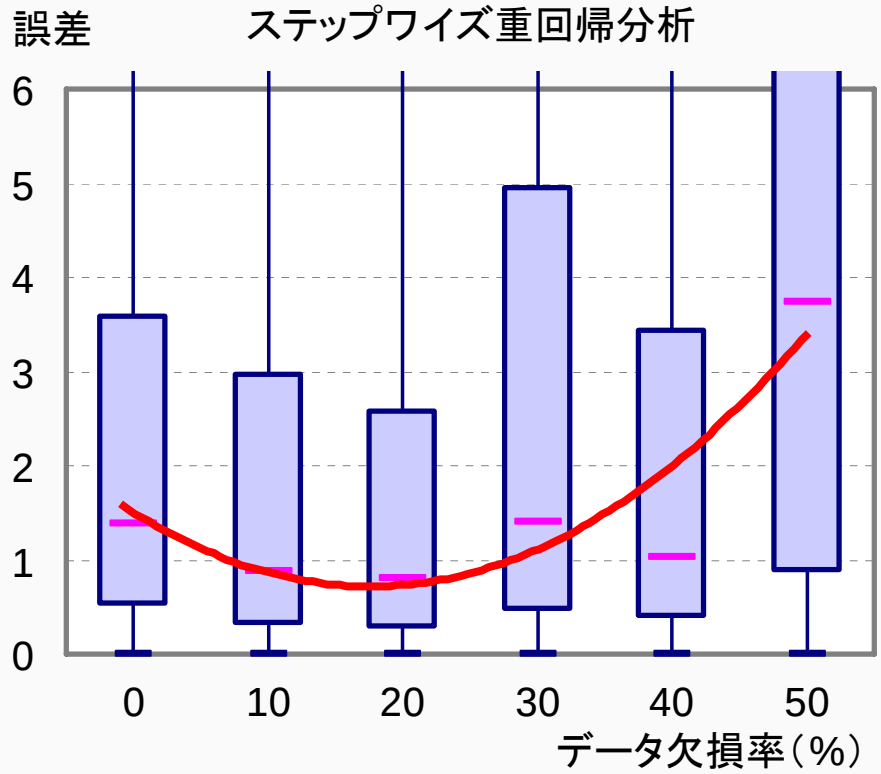
- 日本のソフトウェア開発企業から収集されたデータを用いたケーススタディを実施し，論文などに投稿してきた（詳細については，付録 II の文献を参照）.

	目的	データ	アルゴリズム	結果	組織
[1] - [2]	開発総工数見積もり	特性変数：13 種類 プロジェクト：1081 件 データ欠損率：60 %	類似度計算：コサイン類似度 見積値計算：増幅加重平均（Type-1）	平均誤差：79 % Pred25：37 %	NTTデータ株式会社
[3]	企業横断的データを用いた開発総工数見積もり	特性変数：97 種類 プロジェクト：378 件（15 社から収集） データ欠損率：67 %	類似度計算：調整コサイン類似度 見積値計算：増幅加重平均（Type-2）	平均誤差：64 % Pred25：30 %	独立行政法人情報処理機構ソフトウェア・エンジニアリング・センター
[4]	プロジェクト早期のシステム規模見積もり	特性変数：20 種類 プロジェクト：85 件 データ欠損率：7 %	類似度計算：コサイン類似度 + ユークリッド類似度 見積値計算：増幅加重平均（Type-1）	平均誤差：28 % Pred25：56 %	株式会社日立システムアンドサービス
[5]	コスト超過リスクの予測（判別予測）	特性変数：199 種類 プロジェクト：45 件 データ欠損率：42 %	類似度計算：調整コサイン類似度 見積値計算：単純加重平均	適合率：73 % 再現率：100 % F1値：0.84	株式会社日立製作所

データ欠損率と見積精度の関係 [6]

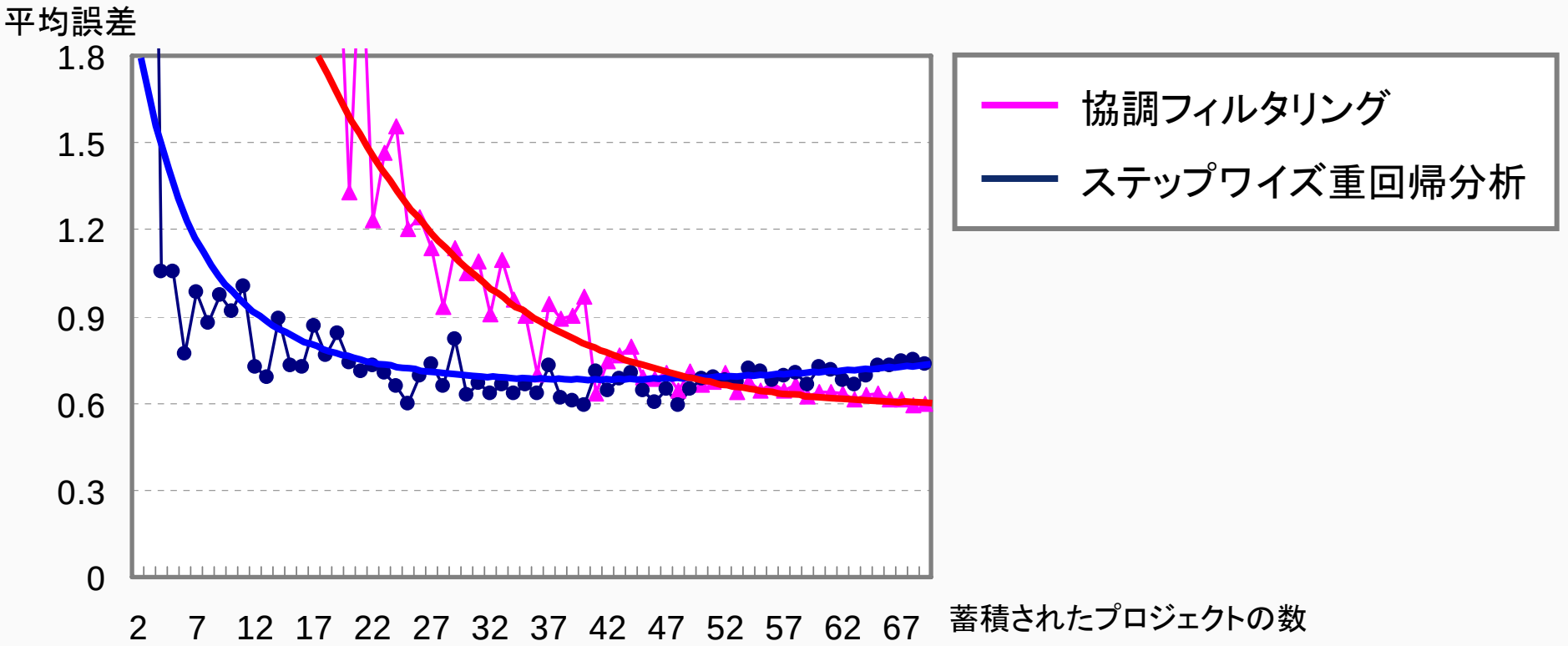
- 目的: 開発総工数見積もり
- データ: 変数10種類，プロジェクト140件，データ欠損率:0-50%
- 類似度計算:コサイン類似度，見積値計算:単純加重平均





プロジェクト数と見積精度の関係 [7]

- 目的: 開発総工数見積もり
- データ: 変数10種類，プロジェクト140件，データ欠損率:0%
- 類似度計算:コサイン類似度，見積値計算:単純加重平均

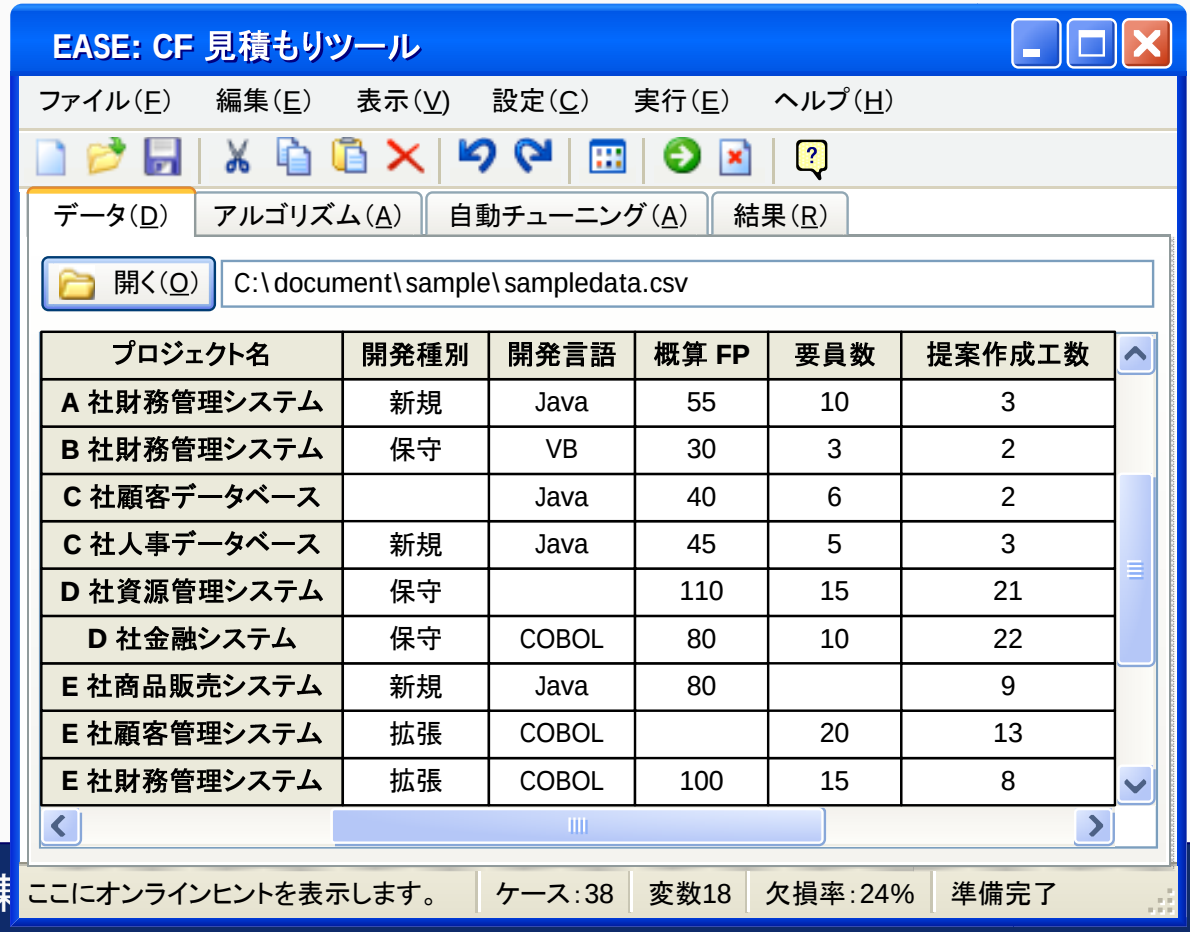


今後の方針

- 手法やツールの普及に努める.
  - 使いやすいユーザインタフェースを備えた GUI ベースの見積もりツールを開発する.
  - 手法やツールの利用を支援するため, ツールの利用説明書やチュートリアルを作成する.
  - 作成した成果物の配布や, 議論のためのフォーラム提供を行う web サイトを作成する.
- 見積精度改善をするために手法／ツールを改良する.
  - 予測アルゴリズムを洗練する.
  - 変数自動選択手法を改良する.
  - データに含まれるノイズを除去する方法を考察する.

特に注力している活動

- 特に GUI ベースの見積ツール開発に注力しています．
  - 機能などについて，御意見，御要望をお聞かせください．
    - 大杉直樹，メール：naoki-o@is.naist.jp，電話：0743-72-5318



プロジェクト名	開発種別	開発言語	概算 FP	要員数	提案作成工数
A 社財務管理システム	新規	Java	55	10	3
B 社財務管理システム	保守	VB	30	3	2
C 社顧客データベース		Java	40	6	2
C 社人事データベース	新規	Java	45	5	3
D 社資源管理システム	保守		110	15	21
D 社金融システム	保守	COBOL	80	10	22
E 社商品販売システム	新規	Java	80		9
E 社顧客管理システム	拡張	COBOL		20	13
E 社財務管理システム	拡張	COBOL	100	15	8

まとめ

- 協調フィルタリングによる見積もり手法を紹介した.
  - データ自動調整機能の実装による受容性.
  - 協調フィルタリングの採用によるロバスト性.
  - 交換可能な予測アルゴリズムによる順応性.
- これまでの適用事例を紹介した.
  - データ欠損率が増加しても，精度は低下しにくい.
  - データが蓄積されるほど，精度は向上してゆく.
- 今後の方針について説明した.
  - GUI ベースの見積ツール開発について御意見，御要望をお聞かせください.

付録 I: 利用可能なツール

- コマンドラインツール
  - 協調フィルタリングエンジン
  - 類似度計算ツール
  - 類似性可視化ツール
  - 類似グループ識別ツール
- Microsoft Excel マクロ
  - 数値見積マクロ
  - 判別予測マクロ
  - 類似性可視化マクロ
  - 自動変数選択マクロ

一部を除き，下記 URL からダウンロードできます．

http://sourceforge.jp/projects/ncfe/

付録 II: 参考文献

[1] N. Ohsugi, M. Tsunoda, A. Monden, and K. Matsumoto, “Effort estimation based on collaborative filtering,” In *Proc. of 5th International Conference on Product Focused Software Process Improvement (Profes2004)*, Lecture Notes in Computer Science, Vol.3009, pp.274-286 (2004).

[2] 角田雅照, 大杉直樹, 門田暁人, 松本健一, 佐藤慎一, “協調フィルタリングを用いたソフトウェア開発工数予測方法,” 情報処理学会論文誌, vol.46, no.5, pp.1155-1164 (2005).

[3] 大杉直樹, 角田雅照, 門田暁人, 松村知子, 松本健一, 菊地奈穂美, “企業横断的収集データに基づくソフトウェア開発プロジェクトの工数見積もり,” SEC journal, No.5, pp.16-25 (2006).

[4] 大杉直樹, 松本健一, 津田道夫, 中屋広樹, 十九川博幸, “協調フィルタリング技術によるソフトウェア規模の予測,” 日立システムジャーナル (2006). (submitted).

[5] 本村拓也, 柿元健, 角田雅照, 大杉直樹, 門田暁人, 松本健一, “協調フィルタリングを用いたプロジェクトコスト超過の予測,” 信学術報, SS2005-39, pp.35-40 (2005).

[6] 柿元健, 角田雅照, 大杉直樹, 門田暁人, 松本健一, “協調フィルタリングに基づく工数見積もりのロバスト性評価,” 日本ソフトウェア科学会FOSE2004, ソフトウェア工学の基礎XI, pp.73-84, (2004).

[7] 柿元健, 角田雅照, 大杉直樹, 門田暁人, 松本健一, “協調フィルタリングによる工数見積もり手法におけるデータ数と見積もり精度の関係の分析,” 日本ソフトウェア科学会FOSE2005, ソフトウェア工学の基礎XI, pp.77-86 (2005).