

使用クラスに基づく ソフトウェアの機能面からの分類

牛窓 朋義 門田 暁人 玉田 春昭 松本 健一

ソフトウェア自動分類への期待

- 特定の機能を持つソフトウェアを探すのに役立つ
- 類似の機能を持つ他のソフトウェアを探すのに役立つ
- ソフトウェア分類学 (Software Taxonomy) の発展に役立つ
- ソフトウェアを蓄積するサイト (SourceForgeなど) における、
利用者のためのデータベースの構築に役立つ

提案手法のキーアイデア

ソフトウェア開発においては、既存のクラスライブラリの中から機能の実現に役立つクラスを選択し、使用することが一般的



ソフトウェアが使用するライブラリクラス群は、そのソフトウェアが実現する機能と強い関連を持つのでは？

提案手法の手順

1. 各ソフトウェアが使用しているライブラリクラス群（使用クラス）を抽出する
2. ソフトウェア間の類似度を示す距離行列を導出する
3. 階層的クラスタリングにより分類する

1. 使用クラスの抽出

- 一般的によく使われているクラスのみを抽出対象とする（例えば開発者の自作クラスなどは対象外）
- 実験ではStigmataを用いた

2. ソフトウェア間の距離行列の導出

tf-idf法を用いて各使用クラスに重み付けを行い、それをもとに距離行列を導出する

tf-idf法

ある単語が**一部の文書のみ**に出現するとき、その単語を**特徴的**とみなして重く重み付けする手法

今回はソフトウェアを文書、使用クラスを単語とみなして適用する

2.1. tf-idf法による重みの導出

ソフトウェア*i*における使用クラス*x*の出現回数

$$tf_{ix} = \frac{C_{ix}}{S_i}$$

ソフトウェア*i*における
使用クラス*x*の出現頻度

ソフトウェア*i*の全使用クラスの出現回数の合計

2.1. tf-idf法による重みの導出

使用クラス x を含む
ソフトウェアの割合の**逆数**

分析対象のソフトウェアの総数

$$idf_x = \log \frac{M}{N_x}$$

全分析対象ソフトウェアのうち、使用クラス x を含むソフトウェアの数

2.1. tf-idf法による重みの導出

$$w_{ix} = tf_{ix} \times idf_{ix}$$

tf-idf法によるソフトウェアiの使用クラスxの重み

2.2. 距離行列の導出

以下の2つを採用した（後の実験で評価する）

- ユークリッド距離
- 重みベクトルのなす角

提案手法の手順(再掲)

1. 各ソフトウェアが使用しているライブラリクラス群(使用クラス)を抽出する
2. ソフトウェア間の類似度を示す距離行列を導出する
3. 階層的クラスタリングにより分類する

評価実験

- tf-idf法による重み付けの効果を明らかにするために、重み付けを**行った**場合と**行わなかった**場合（後述）を比較する
- 距離行列の算出に**ユークリッド距離**を用いた場合と**ベクトルのなす角**を用いた場合を比較する

重み付けなしの場合について

重みなしユークリッド距離は，ソフトウェア i が持つ使用クラスの種類数を UC_i ，ソフトウェア ij 間で共通する使用クラスの種類数を $match_{ij}$ とするとき，ソフトウェア ij 間の距離 d_{ij} は以下の式によって定義される．

$$d_{ij} = 1 - \frac{2 \cdot match_{ij}}{UC_i + UC_j}$$

実験対象

表 1. 分類対象ソフトウェアの内訳

ドメイン	件数
3D First Person Shooters	5
Calendar	5
Chemistry	5
Internet Relay Chat	6
Mail	6
SUDOKU	5

分析対象プロジェクトは全32件

実験結果(1)

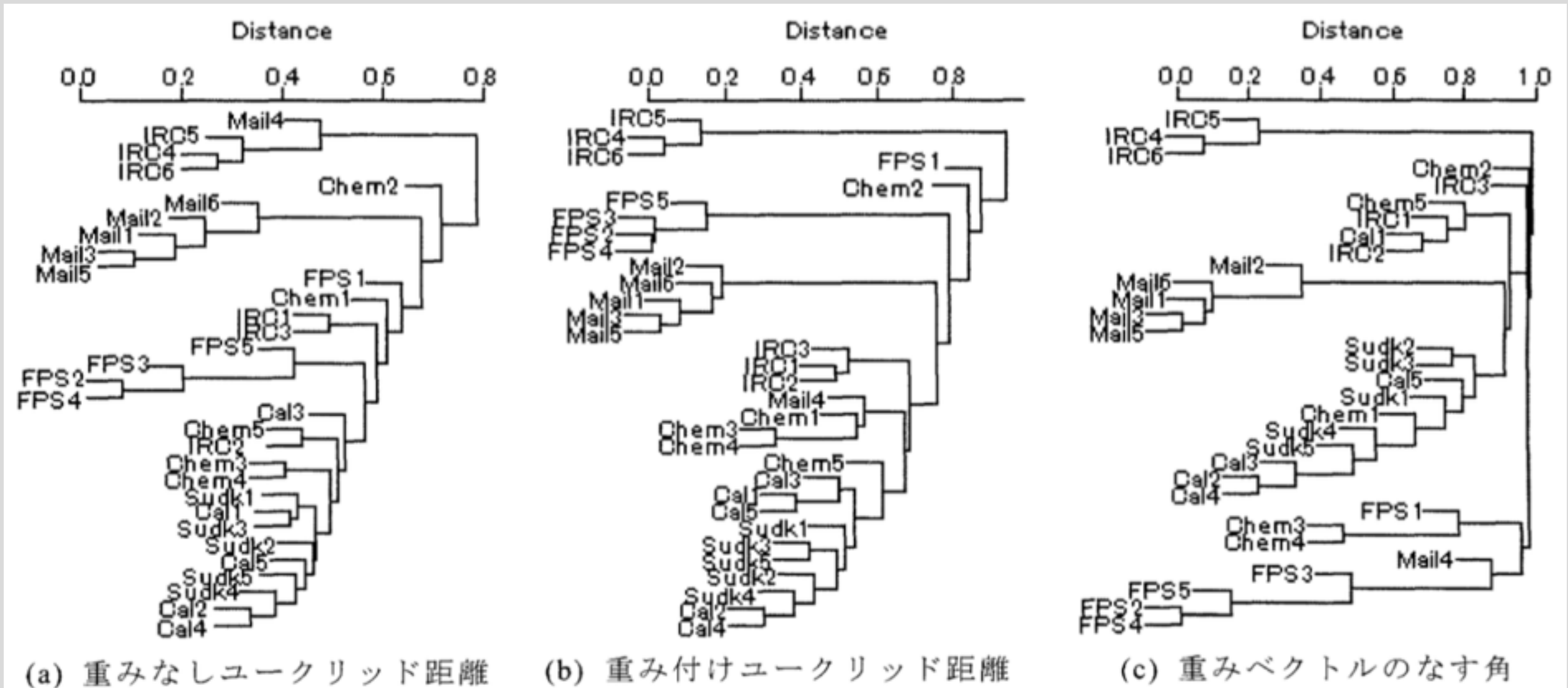
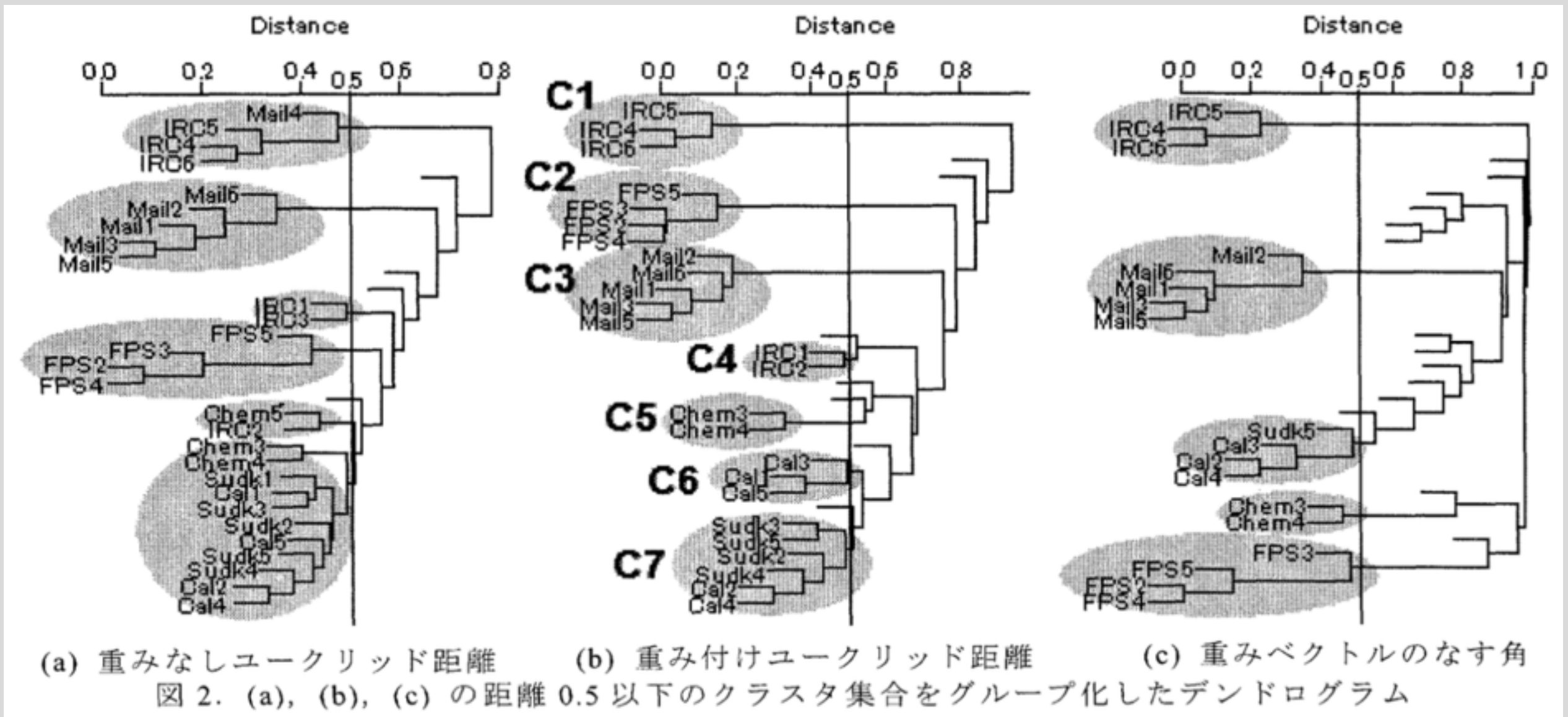


図 1. (a), (b), (c)のデンドログラムの出力結果

実験結果(2)



(b) が最も高い精度で分類できた

考察

- 妥当な分類結果 (C1、C2、C3) が得られたり、重み付けにより精度が向上したもの (C4、C5、C6) があった
- 分類に失敗したもの (C7) もあった
- 分類に失敗したドメインの多くは汎用性の高いクラスばかりを利用しており、ドメイン単位での特徴がなかったと考えられる
- いずれのクラスタグループにも属さなかったソフトウェアは、そのソフトウェアにしか出現しない使用クラスを多数含んでおり、他のソフトウェアとの距離の値が大きくなってしまったものと考えられる

まとめ

- 使用クラスの情報を用いて、多数のプログラムを機能面から分類する手法の提案を行った
- 実験では6ドメイン計32件のソフトウェアを分類し、提案手法の有効性を評価した
- 実験の結果、提案手法ではユークリッド距離を用いる場合に最も妥当な分類結果が得られた