プログラム入門教育における ソースコードの自動分類

情報・通信工学科 学籍番号 1211215 寺田研究室 渡邊裕貴

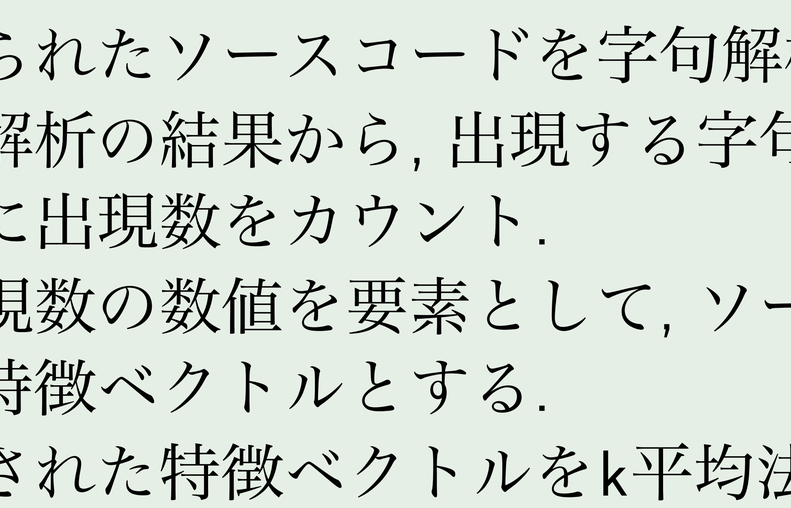
概要

情報工学教育において, あるプログラムを作成するという課題があったとき, 学習者はその課題を解決するにあたって様々なアプローチで実装を行うことになる.
これらのプログラムに対して教員が評価を行う際, ソースコードの数が膨大になり, 学習者がどういったアルゴリズムまたは文法を用いたか等の, 課題への理解度を図る上で重要な要素をは把握することが難しくなると予想される.
そこで本研究では, 集められたソースコードの集合に対してクラスタリングを行い, 似ているコードのグループを特定して, 各グループごとの特徴を明らかにする手法を提案する.

提案手法

1. 集められたソースコードを字句解析する.
2. 字句解析の結果から, 出現する字句の種類ごとに出現数をカウント.
3. 各出現数の数値を要素として, ソースコードの特徴ベクトルとする.
4. 生成された特徴ベクトルをk平均法, 階層化クラスタリングでクラスタリングする.

本研究では特に対象とするソースコードをC言語のソースコードに限定した. カウントされる字句の種類は92種類である. すなわち, 92次元の特徴ベクトルが生成される.



図：特徴ベクトルへの変換

代表要素の選び出し

さらに, 特徴ベクトルの要素から, クラスタリングを行う上で他のベクトルとの差が出やすい要素を見つけ出して, これらの要素のみで構成された特徴ベクトルの集合に対するクラスタリングも行う.
下の表は, ある同じ課題に対する解答として提出されたソースコードの集合を調べ, 最低でも全体で3回出現した上で, 特に出現回数の分散が大きかった字句要素を並べたものである. 本研究では, こういった出現回数の分散が大きい要素が, そのベクトルの特徴を表す代表要素であると判断した.

平均出現数	分散	字句の種類
4.09756	0.0061381212481747905	]
4.09756	0.0061381212481747905	[
3.01219	0.0046903085081204865	(文字列定数)
4.37804	0.003588336533917869	=
5.89024	0.0034536320972965826	*
6.25609	0.0022740175931852983	,
4.32926	0.002145489806431444	}
4.35365	0.0020753447738662656	{
7.36585	0.0019155084387191715	(数値定数)
12.3902	0.0018849712486051786	)

結論

上で述べたソースコードの集合に対して, 実際に92次元の特徴ベクトルと, 代表要素と判断した分散が大きい要素10個のみを含む特徴ベクトルの集合を生成して, k平均法と階層化クラスタリングの両手法を適用した.
いずれの手法においても, ある字句を用いたか用いていないかによって学習者のグループを分離され, おおまかに解答の傾向の似た学習者グループが特定できた.
また, 代表要素のみを選び出したベクトルに対するクラスタリングの実験では, クラスタリングの結果から各クラスタの特徴が捉えやすくなった.