第4章 「関心度」パラメータを追加して分析ライブラリでサッ！

ステップ3：最も精度が高そうな「クラスタ数」を統計的に決める

佐藤 聖

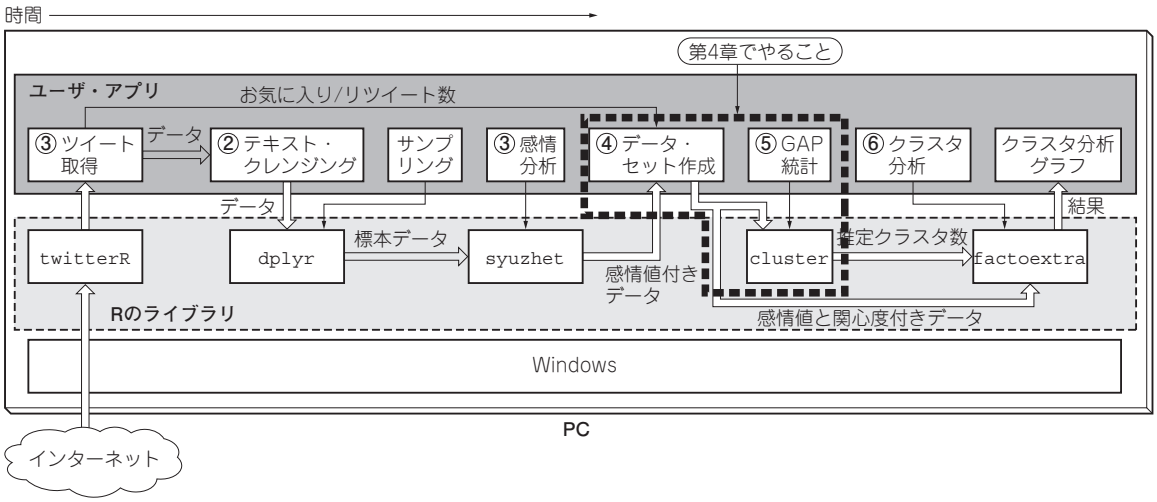

図1　次章での人工知能解析（クラスタ分析）の精度を上げるために…解析時に必要となるデータ群分割数（クラスタ群の数）を統計処理プログラムから求める

今回使う人工知能アルゴリズム「クラスタ分析」（第5章に詳細）を実際に試すには，取得したデータ群を分ける数（クラスタ群の数）を決める必要があります．精度を上げるために，手動ではなくプログラムでクラスタ群の数を求める方法を紹介します（**図1**）．

● 準備…クラスタ分析のライブラリを選ぶ

この後第5章では，クラスタ分析の際に，k平均法というアルゴリズムを用いて，第2章で抽出した「つぶやき」をクラスタに分類します．このk平均法によく利用されるライブラリには，標準ライブラリkmeansがあります．k平均法も行えるクラスタ分析ライブラリclusterもあります．

統計処理向きR言語の統合開発環境RStudioには，どのようなライブラリがあるか検索してみます．**図2**にあるRStudioのHelpタブを開きます．丸で囲った「検索ボックス」に「k-means」と入力して検索すると，k平均法の関連ライブラリを含むパッケージが見つか

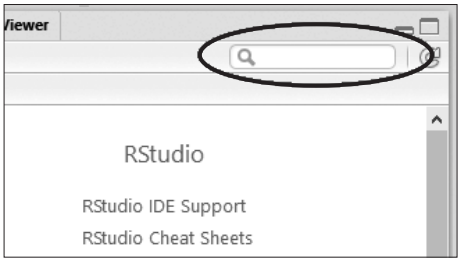

図2　統計処理向きR言語の統合開発環境RStudioでライブラリを検索

表1　主なk平均法ライブラリ付きパッケージ

パッケージ	関　数	説　明
cclust	cclust	凸クラスタリング
e1071	cmeans	ファジーC平均クラスタリング
factoextra	hkmeans	階層k平均クラスタリング
fpc	kmeansruns	クラスタ数（k）と初期設定の推定も行うk平均法
kernlab	kkmeans	カーネルk平均法
RWeka	Cobweb	R/Wekaクラスターズ
trimcluster	trimkmeans	トリミングk平均クラスタリング