研究テーマ

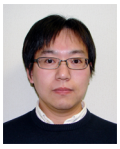
# 与えたデータと類似したデータ集合を得るための クラスタリングシステムの研究

◆キーワード
クラスタリング、意味解析、単語分類

◆産業界の相談に対応できる分野
データマイニング、情報検索、自然言語処理、情報抽出、データ分類、機械学習

工学部情報工学科　講師
佐々木　稔

TEL　0294-38-5159
FAX　0294-38-5152
URL　http://sasa.cis.ibaraki.ac.jp/member/sasaki/
e-mail　msasaki＠mx.ibaraki.ac.jp



一言アピール

本研究は、少量の手がかり情報から関連性の高いデータ群の発見方法を解明するものです。

## 研究概要

あるキーワードについて大量の文書を検索し、抽出した文書集合から、内容についての分類を効果的に行う新しい手法の開発を行っています。内容の似た記事の数が少ない場合は、それを見つけるためのヒントとなる正解データを用意することが難しいため、半教師ありクラスタリングと呼ばれる手法を利用することで、これらの問題を解決し、有効なクラスタリング結果が得られると考えています。そこで、従来手法よりも緩やかな制約条件を与えることで効果的なクラスタリング結果が得られることを目標として、関連のありそうな別のデータを素性として追加し、クラスタリングを行う手法を開発しています。

このクラスタリング手法では、索引語-文書行列に追加する素性として、文書間の類似度行列を追加します(右下図参照)。この類似度行列を追加するために他の文書集合を用意し、元の行列と同様にして索引語-文書行列を計算します。元行列の各列と追加文書行列の各列に対して類似度計算を行い、追加文書を行ベクトルに、元の文書を列ベクトルに対応するような類似度行列を求めます。このようにして求めた類似度行列を元の索引語-文書行列の下に追加し、単語と文書間類似度を要素として持つ合成文書ベクトルを作成します。この合成文書行列に対し、k平均クラスタリングと似たクラスタリング手法を利用してクラスタリングを行います。

このクラスタリング手法の有効性を検証するため、それぞれ120件程度のデータを含む3種類の新聞記事集合を利用して評価実験を行いました。素性を追加しない場合のクラスタリング精度と比較した結果、異なる文書集合の関連度を素性として追加することにより、エラー数が減少し、クラスタリング精度が向上することを確認することができました。また、関連の強い文書集合を追加することにより、クラスタリング精度が効果的に向上することが可能であることも分かりました。

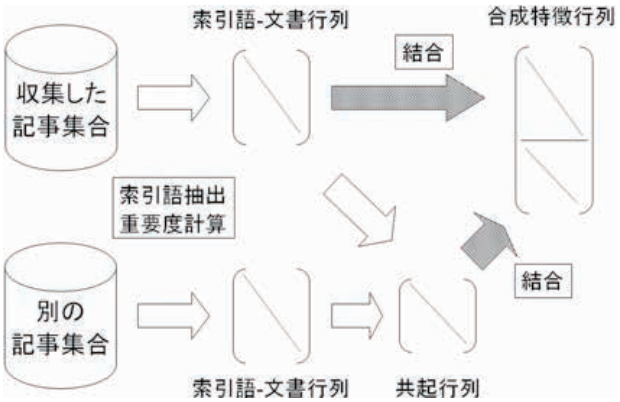



図1 合成特徴行列の作成方法

何に使える?

様々なデータからの特徴抽出、似ているデータの分類などに利用できます。