# 要 旨

## テキストマイニングにおける文書構造を利用した特徴の有効性の検証

滝　優基

文書には，段落や文などの文書構造が存在するが，単語ベクトルによるベクトル空間モデルではこれらを考慮されない．そのため，これまでの機械学習によるテキストマイニングでは，出現単語やその頻度のみが考慮され文脈や文書の構造を考慮しない．そこで，本研究では，文書構造を加味した特徴の，機械学習における有効性の検証を目的として，経済に関するニュース記事から機械学習により株価予測を行い，テキストマイニングへの有用性を確認する．本手法では，文書構造を表現するデータモデルとして文ベクトル集合モデルを用いて文書を表現する．実験データとして，NIKKEI NET，Infoseek と MSN 産経ニュースで公開された経済に関係するニュース記事を 170 件を使用する．交差検証法を用いて，170 記事からランダムで 160 記事を選択して学習データとし，残りの 10 記事を予測データとするデータセットを，学習用の記事を変えて 10 種類作成し，それぞれの学習用の記事に対して 10 回づつ学習と予測の実験を行い，従来手法と比べ 7 ポイント高い認識率となることを示す．また，従来手法では認識率が最大 68%，最低 27%と不安定であったが，本手法では最大 64%，最低 42%と安定した精度となることを確認する．

**キーワード**　文ベクトル，機械学習，テキストマイニング，文書構造，単語ベクトル

# Abstract

## A Study on a Text Feature regarding Document Structure for Text Mining

A document has structures such as paragraphs and sentences, however the vector space model is difficult to represent these text structures. In text mining using machine learning, particularly, the effect of text structures has not been studied sufficiently. The purpose of this research is to clarify the effect of text structure to the machine learning based text mining. In this paper, stock price prediction by the hierarchical neural network and backpropagation are performed to reveal the usefulness of text structures of news articles. The sentence vector set model is used as the data model to represent the document structures. The experimental data are 170 economical news articles in NIKKEI NET, infoseek, and MSN Sankei news. Prediction of stock price movement is performed using cross-validation. The data set consists of 160 training data and 10 trial data from 170 articles. The experimental results show that the recognition rate increases 7 points compared with the conventional method. Also we clarify the recognition rate is more stable using proposed method(the recognition rate spreads 42% to 64%)while the recognition rate is unstable using the conventional method(the recognition rate spreads 27% to 68%)

***key words*** sentence vector, machine learning, text mining, document structure