要 旨

形式概念分析の
ドキュメント・クラスタリングへの応用

森木 彰規

本論文では，形式概念分析のドキュメント・クラスタリングへの応用を提案する．既存のドキュメント・クラスタリングでは，単語ベクトルまたは文書ベクトルとして数値化された情報に対してコサイン距離を用いて数値処理を行うものが主に用いられてきた．しかし，従来の手法では類似するという結論が出ても，実際に文書を読むと内容はあまり類似するとは思えないという結果が出ることがしばしばある．そこで，本研究では形式概念分析を用いることによって，特徴を共有する文書の集合への分類を行う．本論文ではこの形式概念分析を英文のニュース記事文書 100 個または 200 個に対して適用する．そして適用の結果として取得される概念束のコンセプトに着目する．着目するコンセプトから下層に連結するすべてのコンセプトに要素として含まれる文書集合を 1 つのクラスタとすることで，ドキュメント・クラスタリングを行う．クラスタリングの結果，同じ単語を共有する基準で取得されたそれらのクラスタには，それぞれ共通の主題でまとまるものが多いことが分かった．また，上層に位置するコンセプトを選択するほど，下層に分岐するコンセプトのクラスタ群が 1 つに大きくまとめられることが分かった．このように，同じ単語を共有する文書群でクラスタを表すことで，人間にも分かり易い形となる．そのため，共有される単語と文書内容の調査など，分析も容易となる．従来の手法ではクラスタリング結果が数値化された情報であるため，人間にとって分析が困難である．本手法は文書整理や要約に応用できる可能性がある．

キーワード　形式概念分析，コンテクスト，概念束，コンセプト，ドキュメント・クラスタリング，IDF

Abstract

An Application of Formal Concept Analysis to Document Clustering

Akinori Moriki

In this paper, we propose an application of formal concept analysis to document clustering. In conventional clustering methods, numeric data are required. In those methods, numeric processing is performed by cosine distance of numeric data as word or document vector. However, several documents of a cluster is not similar as a result of classification applied conventional methods. In our research, a new clustering method is proposed by an application of formal concept analysis. Documents are classified into sets of documents shared same features by formal concept analysis. In addition, each set of documents can be selected in the method. We, thereby, propose document clustering which is suitable for expressing themes of documents based on information of documents as words. In this paper, formal concept analysis is applied to 100 and 200 documents of news articles of reuters. And then, document clustering is performed by selecting each concept on concept lattice. Elements of each article are included in all concepts connecting to lower layers of a selected concept. Those elements are set as a cluster. Each cluster has a shared topic. In addition, clusters of low-level connecting layers are set as a cluster by selecting concept on higher layers. Such points in our research can be applied to text classification or text summarization.

key words Formal Concept Analysis, Context, Concept Lattice, Concept, Document Clustering, IDF