要 旨

形式概念分析を用いたテキスト分類の検討

森木 彰規

形式概念分析は束論をもとにして提唱されているデータ分析手法である．この手法では分析対象 (オブジェクト)，対象の属性を概念表に表し，そこから Hasse 図を作成し，概念間の半順序関係を表す．そのため数値データで表現することなく視覚的にオブジェクト間の関連性を調査することができる．本研究では，大量のリンクや文書が無秩序に配置された Web ページに対して，形式概念分析を用いてテキスト分類を行うことで，テキストを階層構造に整理し，テキストの要約や主題の推定への応用を検討する．本論文では，英文テキストに対して形式概念分析を実行し，テキスト分類の実験を行う．テキストには英文のニュース記事データベース Reuters–21578 を採用し，各記事をテキストファイルに保存したものを用いる．各テキストファイルをオブジェクト，各テキストの単語を属性としたコンテクストを作成し，形式概念分析支援ソフト「Concept Explorer」を利用して概念束に表す．表示された概念束を調査した結果，英文では必然的に使用される前置詞や冠詞というテキスト内容を示さない単語が概念束の上位層に位置し，多くのテキストで共用されていた．逆にテキスト内容に関わる名詞や一般動詞は概念束の下位層に位置し，2~4 のテキストで共通していた．また今回の実験では，“said” が概念束の上位に位置したが，これは英文のニュース記事では発言者名を記事中に必ず明記するためと考えられ，このことから，概念束の上位に名詞や一般動詞といった重要単語が表れた場合，その単語はテキスト全体の分野やジャンルといった傾向を表すもので，また概念束の下位の重要単語はテキスト間のより詳細な関連性を表すということを確認している．

キーワード　形式概念分析，テキスト分類，概念束

Abstract

A Proposal of Text Classification using Formal Concept Analysis

Akinori Moriki

Formal concept analysis, visualizing relations among objects by partial order relation, is a data analysis method based on lattice theory. The method is used Hasse's diagram which is generated by a 2–dimensional table consisted of objects and attributes. In this thesis, formal concept analysis is applied to articles included in Reuters–21578 for obtaining main subjects and summarization of the articles. Objects are the news articles, and attributes are words included in all of the article. Concept lattice is constructed with Concept Explorer, the software for formal concept analysis. In the result, prepositions and articles, such as "a", "an", and "the", are located on high layer of the concept lattice. This situation indicates that prepositions and articles are common words for many news articles. They are, however, not suggestive for news contents, since those words are not meaningful. On the other hand, nouns and verbs are generally meaningful words, and they are suggestive for news contents. Those words are located on low layers of the concept lattice, and are common for 2 to 4 articles. However, "said" appears on high layer of the concept lattice. This causes by the name of speaker in news. Therefore, nouns and verbs appeared on high layer are indicate tendencies of all articles, and nouns and verbs appeared on low layer indicate relation and association among texts.

key words formal concept analysis, text classification, concept lattice