4P-1

制約のある条件下でのテキストからの有効な情報抽出技術について

津田　和俊†
東北大学大学院情報科学研究科†

１．はじめに

学術論文や新聞記事など、掲載できる文字数に制限の多いメディアでのテキストでは、意図的に必要な情報が省略される場合がある。この場合、本来重要となるキーワードのテキスト中の出現頻度が意図されるものから大幅に変わったり、あるいはテキスト中に存在しない場合がある。

この研究では、このような制約のある条件下でのテキストについて、クラスタリングの手法を発展させ、隠されたキーワードの情報抽出技術について論じる。

２．本研究の対象となるテキストの問題点

例えば学術論文の題名や要旨に着目すると、文字数の制限が厳しく、前後の文脈から人間の目には明らかである情報を省略したり、あるいは同じ項目の繰り返しを極力避けることがある。短い新聞記事などにもこのような例が見られることがある。

このようなテキストを扱う場合、一般的に使われている TF-IDF 法は、本文などのテキストで想定される出現頻度とはかなりかけ離れるものとなる。また、前後の記事の文脈から判別できる情報の省略が行われた場合、テキスト中に意図される必要なキーワードが出現しない場合もある。

このようなケースでのテキストを扱う際に、クラスタリングやキーワード抽出を行うにあたっての困難さがあることを示す。

３．本研究の概要

各テキストの情報を、ベクトル空間モデルに基づき、単語の出現を要素として持つベクトルとして表現する。

この２テキスト間の意味付けを定義する際に、各クラスター内での２単語間に着目した出現頻度の統計量に応じた密度行列を定義し、テキストのベクトルとのノルムを計算することで相関を取る方法を考える。

この計算方法で、テキスト中に含まれる複数の単語のアンサンブルを重み付けすることにより、各クラスターとの相関をスコア化し、テキスト中に隠れた単語の連想を機械的に行う手法を考える。また、他のクラスタリングの技法との差異を議論する。