単語間の距離を利用したクラスター分析による次元圧縮

Dimension Reduction by Cluster Analysis using Semantic Distance between Words

2019Y05　遠藤 大介（Daisuke Endo）
担当教員　出口 利憲（Toshinori Deguchi）・山田 博文（Hirobumi Yamada）

１．序論

近年のコンピュータやスマートフォンの普及は著しく、ソーシャルネットワーキングサービスを利用することが一般的になりつつある。インターネット上で使用されるテキストは増加し続けているが、そのテキストには重要な情報から虚偽の情報まで多くの情報が含まれている。その情報を人間が全て読み判断するには、テキストの量が膨大すぎて困難である。そこでコンピュータを使用してテキストから有用な情報を得るためのコンピュータ技術が開発された。これをテキストマイニングという。テキストマイニングの技術は膨大なテキストデータをデータとして処理することができるが、単語の多義性や書き言葉と話し言葉の違いなど、自然言語から有用な情報を得るには課題も多く現状では完成された技術ではない。

本研究では文書間の類似度計算について、各文書に含まれている単語の意味を考慮した手法を提案し、有効性について検討する。

２．提案手法

本研究では、Word2vec によって求められる単語間の距離とクラスター分析を利用した方法を提案する。

Word2vec によって単語間距離を取得するには、学習済みのモデルが必要である。そこで本研究では、東北大学の乾、岡崎研究室にて作られたモデルを利用する。このモデルは日本語 Wikipedia の本文を元に学習しており、ベクトルは 200 次元で、Skip-Gram 法を採用している。(1)

クラスター分析は、本来は次元削減を行うための技術ではない。しかし、単語間距離を利用して名詞のクラスター分析を行うことで、名詞をいくつかのクラスターに分類することができる。これを利用し、分類した名詞をクラスターごとにまとめることを次元削減とした。この手法をクラスター分析による次元削減とする。

提案手法の有効性を検討する上で、従来から利用されている潜在的意味解析（LSA：Latent Semantic Analysis）と潜在的ディリクレ配分法（LDA：Latent Dirichlet Allocation）を比較対象とする。手法の比較を行う実験には、本校電気情報工学科のシラバスを対象とした。

３．実験方法

始めにシラバスデータの取得を行い、形態素解析を行うことで各シラバスに含まれる名詞を取得した。それらの名詞に対して TfIdf 値の算出を行い、結果を行列に保存した。この行列は文書-名詞行列となっており、各要素には TfIdf 値が格納されている。また、主成分分析を行い、次元削減を行った際の累積寄与率を求めた。その結果、主成分数が 54 の時に累積寄与率が 80%以上になったため、主成分数を 54 とした。その結果、取得した名詞 3287 個を 54 個のクラスターにまとめることができた。

LSA では特異値分解によって得られた単語ベクトル、LDA では各単語が各トピックに属する確率を利用してそれぞれ次元削減を行った。

クラスター分析による次元削減では、まず Word2vec を利用して名詞間の単語間距離を算出した。次に、クラスター分析を用いて次元削減用のクラスター-名詞行列を作成した。クラスタリングにおいては、クラスター間の距離の測定方法にはウォード法を採用した。また、行列を作成する式には、昨年までの研究で使用されていた下記の式(2)を利用する。ここで、$a_{i,j}$はクラスター-名詞行列に格納する要素、C_iは各クラスター、$|C_i|$はクラスター内の要素数、$S_{k,j}$は 2 つの単語w_k、w_j間の類似度、w_jは単語を示す。

$$a_{i,j} = \begin{cases} \frac{1}{|C_i|} \sum_{k \in C_i} S_{k,j} & (w_j \in C_i) \\ 0 & (w_j \notin C_i) \end{cases} \quad (1)$$

最終的に、次元削減用行列を文書-名詞行列に掛け

合わせ、次元削減を行った。

LSA、LDA、クラスター分析による次元削減のそれぞれにおいて、cos 類似度を用いてシラバス間の類似度を算出した。そして、類似度を計算した科目間のつながりを可視化するためにデンドログラムの作成を行った。

４．実験結果

LSA、LDA、クラスター分析による次元削減の各デンドログラムについて、特徴等を以下に示す。

LSA のデンドログラムは、科目を大きく 3 つに分類しているという特徴が見られる。各分類は実験や研究の分類、一般科目の分類、専門科目の分類といった形に分かれている。しかし、これらの分類はデンドログラムの全体像を観察した際にでてくる特色である。そのため、各分類に属する科目を細かく確認すると、上記した分類とはかけ離れた科目が属している場合がある。

LDA のデンドログラムでは、科目が結合するタイミングが早く、大きく分けて 2 つの分類となっている。しかし、科目名からはこれらの分類の基準や傾向が読み取れず、科目の分類は上手く行われていないように見える。また、早い段階での結合を観察すると、電気系、情報系、一般科目で結合しやすい傾向にはあるように見える。

クラスター分析による次元削減のデンドログラムでは、大きく分けて専門科目と一般科目の 2 分類となっている。また、専門科目の分類のなかでも、情報系と電気系の科目にクラスターが 2 つ分かれている。デンドログラムより専門科目の分類の様相がよくわかる部分を抜粋した結果を Fig. 1 に示す。

５．考察

上記の実験結果から、クラスター分析による次元削減は LSA、LDA よりも上手く文書の分類ができているとわかる。

しかし、Word2vec を用いて名詞間の単語間距離を算出する際、取得した名詞 3497 個の内 210 個の名詞が学習済みモデルに登録されていなかった。これらについてはクラスター分析による次元削減においては考慮されていない。今回は、除外された名詞が単語間距離算出に利用した単語よりも圧倒的に少なく、割合にして 6%ほどである。そのため、類似度計算の結果に影響はあまり出なかったと考えられるが、もしこの割合がもっと大きくなった場合は、本提案手法の有効性は確実に失われていく。そのため、単語が単語間距離算出に利用しているモデルに登録されていなかった場合は、なんらかの処理を施して類似度計算への影響を出来るだけ少なくする必要がある。

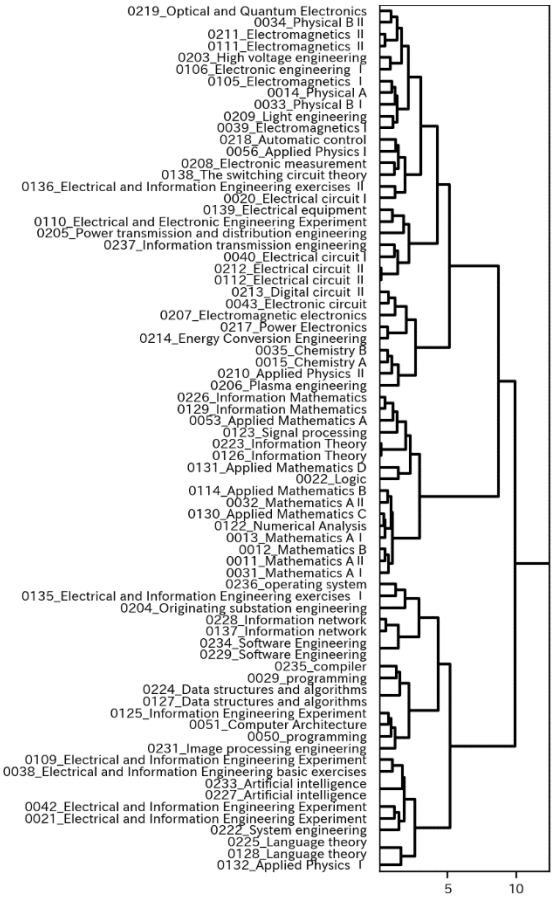

Fig. 1 Part of the result with dimension reduction by cluster analysis

６．結論

今回の中間発表までの結果だけであれば、本提案手法には既存手法には無い利点や、有効性を見つけることができた。また、単語間距離算出手段に日本語 WordNet を利用した場合と比較し、単語間距離算出手段が的確であるのかという点も今後の課題となる。

参考文献

(1) 鈴木正敏，日本語 Wikipedia エンティティベクトル. http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/ （閲覧日：2020 年 6 月 5 日）

(2) 服部修平，単語間の概念距離を用いたクラスター分析による次元圧縮，岐阜工業高等専門学校電気情報工学科特別研究報告，平成 30 年度.