

概念距離を利用したクラスター分析による次元圧縮

Dimensionality Reduction by Cluster Analysis using Conceptual Distance

2017Y26 服部 修平 (Shuhei Hattori)

担当教員 出口 利憲 (Toshinori Deguchi), 田島 孝治 (Koji Tajima)

1. はじめに

テキストマイニングによる文書間の類似度の計算方法として、概念距離を利用したクラスター分析による次元圧縮から類似度の計算ができると考えた。この手法の有用性を求めるために、文書館の類似度計算に用いられる次元圧縮の手法の LSA(潜在的意味解析)と比較する実験を行なった。実験の結果から本手法の改善ができることが分かった。改善を目的に、圧縮する次元数を増やすことで概念距離を利用したクラスター分析による次元圧縮の類似度計算の変化を調べた。

2. 計算方法

日本語 WordNet⁽¹⁾ から概念距離を求める方法を示す。単語 a, b があるとき、ルート synset から最も遠く単語 a, b で共通の synset の段数を C_{ab} 、共通の synset から a, b への最短距離を N_a, N_b とすると、概念の類似度 $S_{a,b}$ は次式で表せる。⁽²⁾

$$S_{a,b} = \frac{2C_{ab}}{N_a + N_b + 2C_{ab}} \quad (1)$$

式 (1) は最大値が 1 になるよう正規化されているので、 $D_{a,b} = 1 - S_{a,b}$ で概念距離 $D_{a,b}$ に変換できる。

次に概念距離を利用したクラスター分析による次元圧縮を用いた文書間の類似度の計算方法について示す。この方法は、名詞の概念距離をクラスター分析することによって分類し、概念を利用した次元圧縮とするものである。

式 (1) で求めた概念距離をワード法によりクラスター分析し、単語-クラスター行列を求める。重要度を Tf-Idf とする文書-単語行列と、単語-クラスター行列を掛け合わせるにより次元圧縮とし、文書-クラスター行列を求める。文書-クラスター行列を cos 類似度で文書-文書行列としたものを文書間の類似度とし、ワード法でクラスター分析し樹形図を求める。単語-クラスター行列の値 $a_{i,j}$ は、各クラスターを $C_i (i = 1, 2, \dots, N)$ 、単語を j とすると、次式 (2) で求められる。

$$a_{i,j} = \begin{cases} \frac{1}{|C_i|} \sum_{k \in C_i} S_{k,j} & (j \in C_i) \\ 0 & (j \notin C_i) \end{cases} \quad (2)$$

Table 1 Group comparison in cluster dendrograms

グループの名前	科目数	本手法の結果
電気磁気学	3	一致
電気材料	2	一致
電気回路	3	別科目が一つ増えた
実験・研究	7	科目が1つ減った
プログラミング	6	3つ以上に別れた
応用物理	4	2つに別れた
光工学	3	科目が1つ減った
電気機器	3	2つに別れた
デジタル回路	5	3つ以上に別れた

3. 実験 1：シラバスの分類による比較

文書として、岐阜高専の平成 28 年度電気情報工学科情報コースの 58 科目のシラバスを対象とする。シラバスを対象とする理由として科目間の類似度から学習に対して有用な知識が得られると考えた。シラバスに含まれていて日本語 WordNet の辞書に登録されている名詞は 988 語だった。LSA を利用した文書間の類似度を比較対象とすることにした。LSA の重要度として Tf-Idf を利用する。圧縮する次元は累積寄与率が 80 % を超えた 29 次元とした。次元圧縮として LSA を行い、その cos 類似度をワード法でクラスター分析し樹形図を求めた。樹形図から 36 科目を類似していると判断し 9 つのグループに分類した。概念距離を利用したクラスター分析による次元圧縮でも同様に樹形図を求めた。比較として、LSA で求めた 9 つのグループに含まれる科目に注目して比較を行った。

4. 実験 1 の結果

本手法による樹形図と、9 つのグループを比較した。結果を Table 1 に示す。グループと違う分類となった部分木は科目名からは類似性が確認できなかった。結果として、14 科目は LSA と同じように分類ができたと判断した。しかし、類似性がない科目でできた部分木は LSA より多く、科目数も多かった。また、分類できていない科目が集まった部分木が存在することがわかった。本手法の結果が LSA と比較してもあまりよくないことがわ

