編集距離を利用する文書クラスタリング

吉原　堅斗　　三浦　孝夫
法政大学理工学部創生科学科

1. 前書き

本研究は、文書の類似性を編集距離で扱い、文書類似性を反映した新しいクラスタリングを提案する。また、類似文書は、初期値などに関わらず同じクラスタに含まれると期待でき、本稿ではクラスタリングを「凝集度」により評価する。

2. 文書クラスタリングと編集距離

従来の文書クラスタリングは、文書のベクトル化を前提とし、単語や文章の並びを考慮しない。しかし、実際には文書の意味が並びに依存するため、内容を正確に理解し類似性を比較することは難しい。この問題を、文字列の近似距離を求める編集距離を活用し解決する。本研究では、文字列自体を単文、文字を単語とみなすことで、語順を考慮した単文同士の類似を表すことができる。

3. 実験

(1)実験準備

コーパスは、2017 年 1 月 3 日～9 日毎日新聞記事 249 記事。頻出する単語で潜在意味解析を行い、類義語は同一単語とする。

(2)提案手法

本研究では、文書のクラスタリングにその類似性を編集距離で扱う手法を提案する。文書は、複数文章から構成され、文章とその並びによって意味が決定する。その文章もまた単文とその並びによって、単文もまた単語とその並びによって意味が決定する。

ならば文書の類似性を編集距離で扱うには、文書の編集距離の編集距離(文章の編集距離)の編集距離(単文の編集距離)で求まる。

評価は、クラスタの凝集度で行う。凝集度を評価することで、編集距離を扱うことが、クラスタとしてまとまりやすい類似性になっているかが分かる。

凝集度 k(s)=Σs^2 D(s)　　(s=1,…n)
　　D={d(1)…d(t)}　　(t=1,…n)
※d はクラスタＤの要素で、現在と前のクラスタで t 個同じクラスタに存在する。

①編集距離を用いる手法で k-means 法
②語頻度ベクトルを k-means 法
①、②の手法でクラスタリングを 3 回ずつ行い、凝集度を求める。

(3)実験結果

表 1　凝集度の平均

	①	②
[1.2]回目	2819	1608.5
[1.3]回目	2463	1482.5
[2.3]回目	2923	1646
平均	2735	1579

4. 考察

表1より、①の編集距離を用いる手法の凝集度は 2735 で②の語ベクトルの凝集度 1579 より高い。このことから文書の類似性に編集距離を用いることは,クラスタとしてまとまりやすく、文書の潜在的な意味をよりは明白に区別している手法といえる。

5. 結論

文書の類似性に編集距離を用いることで語順を考慮することができ、より類似記事同士がまとまったクラスタリングができる。

6. 参考文献

[1]山口　信：変動する検索語の近似文字列検索, DEIM, pp. 1-3, 2014.