6J-5

ドキュメント分類のための単語クラスタリング手法の提案

稲吉洋† 岡本秀輔†

†成蹊大学大学院理工学研究科

概要

本稿では、ドキュメント分類を目的とした単語のクラスタリング手法の提案を行う。本研究の目的は、ドキュメント単位でのクラスタリングを行い膨大なドキュメントファイルについて正しく分類し、その関係性を把握出来るようにすることである。そのための手法として、1個のドキュメント及びドキュメント集合のレベルでそれぞれ単語をクラスタリングを行う手法を考えた。その基礎実験として、検索エンジンを用いた単語のクラスタリングについて報告する。

1 研究背景

コンピュータやWebの技術の発達に伴い、扱う情報量も膨大なものとなってきている。その一例として、技術文書がPDFファイルとして容易に入手出来るようなサービスがある。それらにアクセスするためには、大量のファイルを自動で分類出来るようなシステムが必要がある。本研究では、効率のよいテキストマイニングを目的とした単語レベルでのクラスタリング手法を提案する。研究の方針として、以下の3点を設定する。

- 1個のドキュメント内にて登場する単語のクラスタリングを行い、ドキュメントの特徴点 (複数個の単語) を定義する。
- 特徴点を定義したドキュメントの集合において、それらの特徴点を用いて再度クラスタリングを行う。
- 辞書 (コーパスなど) を事前に用意する必要がないシステムを構築する。

本稿では、上記の方針のもとで提案するアルゴリズムについての詳細と、予備実験として事前に用意した単語の集合においてクラスタリングを実行した際の評価について述べる。

The Word Clustering Method for Classifying Document Files
Yo INAYOSHI†, Shusuke OKAMOTO†
†Department of Information Science, Seikei University
3-3-1,Kichijhoji,Musashinoshi,Tokyo,180-8633 Japan

dm106202@cc.seikei.ac.jp, okam@st.seikei.ac.jp

2 アルゴリズム

クラスタリング手法としてk平均法 [1] を用いる。k平均法は非階層クラスタリング手法の1つであり、多数のデータについて簡易な計算によって分類することが可能である。本節でも、k平均法を用いる。

クラスタリング対象となるn個の単語の集合 $W = w_0, w_1, ..., w_n$ について、2個の単語 w_i, w_j の学術論文検索サービス Google Scholar [2] における AND 検索のヒット数 h_{ij} を取得する。h_{ij} の値が大きいほど w_i と w_j が共起する確率が高く、相関性が強いことが言える。これを元に、係数 r_{ij} を以下のように定義する。

$$r_{ij} = log_2(\frac{10^m}{h_{ij}}), \quad m = max(log_{10}(h_{ij}))$$

この式は、エントロピ計算式を基にしており w_i と w_j が共起する確率が高いほど、r_{ij} の値は小さくなる。これらを用いて、図1のような対象行列を作成する。行列はそれぞれ行および列の総和について降順に整列し直す。ここで、行および列の総和が大きいものを上位、小さいものを下位と定義する。単語集合全体をk個のクラスタに分類することを考えたときに、計算によって暫定的に定義されるクラスタを $C_g(g = 0, 1, ..., k)$(要素数 $member_g$)、C_g の重心の座標 $cent_g$ および w_i との相対距離 w_{gi} は以下に定義する。

$$cent_g = (\sum_{j=0}^{n-1} m_{j0} * \frac{1}{member_g}, ..., \sum_{j=0}^{n-1} m_{j(n-1)} * \frac{1}{member_g})$$

$$d_{gi} = \sqrt{\sum_{j=0}^{n} (r_{ji} - cent_{gi})^2}$$

生成するクラスタ数をkとし、上位の単語から順にクラスタ $C_{mod(i,k)}$ に割り当てる。これにより各単語の意味的な相対距離をn次元空間座標で表す。w_i にとって d_{gi} が最小となるとき、gを w_i の所属クラスタとする。以降、上記の計算を繰り返し行い、各クラスタの構成要素に変化がなくなり次第処理を終えて C_g を確定させる。

クラスタの内容	a	b	c	正解率
哲学 文学 心理学 社会学 経済学 教育学 法学 政治学 言語学 史学 経営学 文化人類学 数学 物理学 天文学 建築学 人類学 農学 薬学：C_0	12	4	3	63.16
歯学 看護学 生物科学 水産学 農芸化学 農業経済学 基礎医学 基礎生物学 畜産学・獣医学 境界農学 農業工学 境界医学 社会医学 内科系臨床医学 外科系臨床医学：C_1	0	0	15	100.00
機械工学 土木工学 人文地理学 地球惑星科学 材料工学 電気電子工学 プロセス工学 基礎化学 総合工学 材料化学 応用物理学・工学基礎 プラズマ科学 複合化学 森林学：C_2	1	12	1	92.31
全体集合 (48 個)				81.25

表 1: 3 分割の結果 (a:人文社会、b:理工、c:生物)

クラスタの内容	a	b	c	系一致
哲学 文学 心理学 社会学 経済学 教育学 法学 数学 物理学 農学 薬学：C_0	7	2	2	a:63.64
政治学 言語学 史学 経営学 文化人類学 天文学 人類学：C_1	5	1	1	a:71.43
歯学 水産学 農業経済学：C_2	0	0	3	c:100.0
看護学 基礎医学 社会医学：C_3	0	0	3	c:100.0
人文地理学 総合工学：C_4	1	1	0	-
機械工学 建築学 土木工学 材料工学 電気電子工学 基礎化学：C_5	0	6	0	b:100.0
地球惑星科学 材料化学 生物科学 農芸化学 農業工学 基礎生物学：C_6	0	2	4	c:66.66
プロセス工学 応用物理学・工学基礎 プラズマ科学 複合化学 畜産学・獣医学 境界農学 森林学 内科系臨床医学 外科系臨床医学 境界医学：C_7	0	4	6	b:40.00 c:60.00

表 2: 8 分割の結果 (a:人文社会、b:理工、c:生物)

3 評価実験

3.1 実験概要

前節で述べたアルゴリズムを用いて、実際にクラスタリングを行う。クラスタリング対象として、科学研究費補助金 (文部科学省が制定) の細目表 [3] にあげられた用語を用いる。細目表のうち、3 系統 8 分科の 48 分野の名称 (総合領域、複合新領域を除く) を用いる。

上記の単語について、Google Scholar にて検索を行い、その結果を h_{ij} としてクラスタリングを行った。今回は、細目表に沿う形でクラスタ数を 3 及び 8 とした。

3.2 実験結果

実際にクラスタリングを行った結果を表1、2 に示す。

表 1 は、単語集合について 3 分割を行った際の結果を示している。3 分割を行った場合は単語全体の 81.25%について正確に分類し、各クラスタ内での正解率についても C_1、C_2 では、合わせても 2 個の不正解に止まった。この実験においては、ほぼ問題なく単語の分類を行うことが出来たと言える。

続いて、8 分割を行った結果について述べる。表 2 の 4 列目の「系一致」は、クラスタ内の登場単語の中で同一の系に属する単語の確率を示す。8 分割の結果について、分科単位で見ると細目表の結果にはそぐわないものとなったが、その中でも同一の系同士で固まっている傾向にあった。両実験の結果より、各単語の大局的な傾向を定義することが出来ると言える。8 分割の中での系単位での分布について表 2 にまとめた。医学系、工学系、農学系の単語について特に同一クラスタに分類される傾向が強かった。

両実験に共通して言えることとしては、行列作成時に単語の並べ替えを行った際に上位の単語が含まれるクラスタについて分類の精度が落ちる傾向が見られた。上位の単語の行や列の数値が重心計算の際にその数値に影響を与えているものと考えられる。

4 今後の展望

今回提案した手法によって、単語の特徴を定義することが出来、テキストマイニングの手法としての可能性を確認した。今後は、同ようのアルゴリズムを 1 個のドキュメント内で実行し、ドキュメントそのものの特徴点を定義することを考える。1 個のドキュメントもドキュメントの集合も、無数の単語の集合であることに違いはなく、同一のクラスタリング手法を適用することで統一されたクラスタリングを実現出来るものと考える。k 平均法そのものは、排他的なクラスタリングを行うものだが、冒頭で述べたようなドキュメントの特徴点を定義することにより、柔軟なクラスタリングが実現可能であると思われる。問題点として、初期値の設定によってクラスタリングが変動する可能性があるため、結果を安定させるような定式化を定義する必要があり、今後の課題となる。

参考文献

[1] マイケル・J・A・ベリー著, SAS インスティチュートジャパン, 江原淳, 佐藤栄作訳： データマイニング手法, 海文堂

[2] Google Scholar, http://scholar.google.co.jp/schhp?hl=ja

[3] 文部科学省, http://www.mext.go.jp/a_menu/shinkou/hojyo/07030710/002.pdf,
科学研究費補助金「系・分野・分科・細目表」の改正について