文書クラスタの判別のための特徴表現付与

小川 知也, 落谷 亮, 西野 文人
(株) 富士通研究所
{tomy,ochi,nisino}@flab.fujitsu.co.jp

1 はじめに

インターネットの普及や全文データベースの増加に伴い, 我々の身の回りの文書情報は飛躍的に増大している. それらの文書に含まれる情報を効率的に活用するために, 文書情報を分類して整理することが行われる. その際, それぞれの文書クラスタに対してそのクラスタを特徴付ける分かり易いキーワードを付与しておくことが望ましい.

クラスタリングに対する特徴表現付与としては, クラスタの特徴ベクトルなどとの類似度の高いいくつかの単語を並べるという手法がある. しかし, 例えば分類した文書集合のフォルダに名前を付けたり学会大会のプログラムの作成などの応用を考えると, 他のクラスタとの相違を際立たせた一つあるいは少数の表現でクラスタを特徴付けることが望まれる. そこで我々は, 文書の内容を反映するような情報を含む複合語に注目し, 複合語を含む特徴表現の付与手法を提案する.

2 文書クラスタへの特徴表現付与

2.1 クラスタリング手法

文書を対象としたクラスタリング手法には階層的クラスタリング手法 [1] がよく知られ, 他にもターム共起に注目した手法 [2], [3] などがあるが, ここでは文書とタームが同じ空間上にマッピングされるという性質に注目し, LSI(Latent Semantic Indexing)[4] に基づくターム / 文書クラスタリング手法を用いる.

この手法を簡単に説明すると, まず LSI に基づき, 文書セットから作成したターム - 文書行列を特異値分解することで次元圧縮されたターム / 文書空間を得る. この空間における各次元軸と文書, タームベクトルとの類似度 (cos) がある閾値以内の文書, タームをそれぞれクラスタにまとめる. 次元軸の + 側, − 側のどちらかに文書やタームが偏る場合は, 分布の小さい方はクラスタとして抽出しない. また, 文書クラスタの全体あるいは上位一部が他のクラスタとの重なりが大きい場合, そのクラスタは他方のクラスタとの関連が高いとして, 抽出しない.

この手法の主な特徴は次の通りである.

- 同じ基準で文書とタームを分類することが出来るため, 文書クラスタに対応するタームが同時に得られる.
- クラスタ間のオーバーラップありの分類を行うため, 複数の特徴を持つ文書やタームは複数のクラスタに分類される. もし文書を一つのクラスタにだけ分類したければ, 最も次元軸類似度の高いクラスタなどに分類すればよい.
- 特異値分解の解法に Lanczos 法 [5] を用いることにより, 処理時間はターム - 文書行列の非零要素数に比例する程度で抑えられる.

2.2 タームの選定

タームをどのように選定するかはクラスタリング精度や処理時間に大きく影響する [6]. 複合語のような長単位のタームを使う方が詳細な分類が可能であるが, 文書数が少ない場合にはスパース問題が起こり, 良いクラスタリング結果が得られない. そこで我々は, 名詞である単語のみをタームとする場合, 名詞単語 + 名詞句 (名詞の最長連続), 任意の名詞部分列の三種類のターム選定を使って実験を行った.

名詞単語 + 名詞句や名詞部分列では独立ではないタームを多数加えることになり, 個々のタームに直交性を仮定するクラスタリング手法ではタームや文

書間関連度などの計算の際にそれらのタームが強く効き過ぎクラスタリング精度に悪影響を及ぼす恐れもある．しかし本稿で用いる LSI に基づくクラスタリング手法では，共起の度合いの高いターム同士は直交性が低くなるように次元圧縮されるため，重みなどを適切に設定すればクラスタリング精度への悪影響はあまり大きくないと思われる．

2.3 タームの重み

名詞単語のタームの重みとしては名詞単語の頻度を用いる．ターム - 文書行列の要素はこの重みに対しさらに IDF や 文書長による正規化による重み付けを行ったものとなる．

名詞単語と名詞句や名詞部分列との間の影響力のバランスをとるために，名詞列タームの重みを名詞列タームの頻度に設定した．ターム - 文書行列の要素は，この重みに IDF や 文書長による正規化による重み付けを行ったものとなる．

2.4 特徴表現付与

出現頻度が低く特徴表現として適切ではないものが付与されるのを防ぐために，文書クラスタに対応するタームクラスタのタームのうち，座標がある値以上のタームを特徴表現とする．座標を絞り込みの基準に用いるのは，上記のようなタームの重み付けの下での座標はクラスタにおけるタームの出現頻度状況を反映すると考えられるためである．この中から，後述する特徴表現の整理により特徴表現を絞る．

3 実験と考察

3.1 文書セットとターム

実験に用いる文書セットとしては，情報処理学会自然言語処理研究会 (NL 研) のホームページ[1] の論文のうち，概要付きの第 114 回から第 128 回を対象とし，タイトル＋概要を文書内容とした．これらのうち，英語を多く含む論文 15 件を除いた 287 件を対象とした．

名詞列をタームに加えることでターム数が増加する．そのターム増加による処理効率低下を抑えるため，文書頻度 2 以上のタームに絞る場合も実験した．それぞれの場合のターム数と非零要素数を表 1 に示す．ターム数の括弧内は名詞を 1 とした場合の比を表す．ターム数の増加は記憶容量の増加につながり，非零要素の増加は処理時間の増加につながるが，名詞列をタームに加えたことによるターム増加はかなり低めであり，特に文書頻度 2 以上のタームに絞った場合はターム数などの増加が低く抑えられている．その理由として，今回の文書セットは文書長が比較的短いこと，文書セットの内容が自然言語処理に限定されており内容のばらつきがあまり大きくないことなどが挙げられる．定量的な評価は行っていないが，文書頻度 2 以上のタームに絞っても分類精度に大きな低下は見られないようである．従って以降では文書頻度が 2 以上のタームに絞って実験を行う．

タームの種類	文書頻度	ターム数	非零要素数
名詞	1 以上	1313 (1.00)	4198 (1.00)
名詞＋名詞句	1 以上	2252 (1.72)	5322 (1.27)
名詞部分列	1 以上	2863 (2.18)	6098 (1.45)
名詞	2 以上	550 (1.00)	3435 (1.00)
名詞＋名詞句	2 以上	661 (1.20)	3731 (1.09)
名詞部分列	2 以上	750 (1.36)	3985 (1.16)

表 1: 文書セットの特性値

3.2 特徴表現付与の実行例

タームが名詞単語の場合の特徴表現付与例を図 2 に，名詞単語 + 名詞句の場合を図 3 に，名詞部分列の場合を図 4 に，それぞれ示す．タームに名詞列を加えることで元の文書セットの内容をより多く反映するような特徴表現が付与されることが分かる．

3.3 特徴表現の整理

タームに名詞列を加えたことで，

> 解析, 形態素 - 解析, 形態素, 日本語 - 形態素, 日本語 - 形態素 - 解析, ...

と似た特徴表現が付与されてしまう．そこで，タームの名詞列階層に注目した特徴表現の整理を行う．

ターム t_i が名詞列としてターム t_j に含まれる場合，t_i は t_j の上位語 (より一般的な概念を表す語)，t_j は t_i の下位語 (より特殊化された概念を表す語) と考えられる．例えば「解析」は「形態素 - 解析」の上位語といえる．もし「解析」の出現頻度状況が「形態素 - 解析」とほとんど変わらないならば，「解

[1] http://cactus.aist-nara.ac.jp/staff/utsuro/SIGNL/

析」は「形態素 - 解析」という形で用いられることが多く，上位語である「解析」は特徴表現から省くことができよう．そこで，ほぼ等しい座標を持つ下位語を持つタームを特徴表現から省くことで特徴表現の整理を行う．

例として，形態素解析関連のクラスタについて考える．タームとして名詞部分列を用いる場合の，各タームの次元軸類似度と座標を図 1 に示す．

類似度	座標	ターム
0.65	0.33	解析
0.60	0.29	形態素 - 解析
0.60	0.30	形態素
0.56	0.12	日本語 - 形態素
0.56	0.11	日本語 - 形態素 - 解析
0.46	0.15	確率 - モデル
0.44	0.20	確率
0.37	0.22	モデル
0.27	0.14	日本語
0.25	0.11	構造

図 1: 形態素解析関連のタームクラスタ

「解析」および「形態素」は「形態素 - 解析」があるので省略できる．一方，「形態素 - 解析」と「日本語 - 形態素 - 解析」では，「形態素 - 解析」を省くほどは両者の座標は近くない．この場合「日本語 - 形態素 - 解析」はこのクラスタの特徴の一面を表しているだけと考えられるので，特徴表現から省くことも考えられる (後述の例では括弧付きで残している)．

タームとして名詞句を含める場合の各名詞句の座標はタームとして名詞部分列を用いる場合のそれに比べ等しいか小さくなることが多く，特徴表現の整理という観点からは，名詞句を含める場合の座標よりも名詞部分列を用いる場合の座標の方が適切な場合が多い．タームとして名詞部分列を用いた時のターム増加があまり大きくないこと，クラスタリング精度の低下も特に見られないことを考えると，特徴表現付与にはタームとして名詞部分列を用いるのが有効と考えられる．

タームとして名詞部分列を用いた場合の特徴表現整理後の各クラスタへの特徴表現付与および文書クラスタの例を図 5 に示す．

4 おわりに

文書クラスタの内容をよく反映するような特徴表現付与手法の提案を行った．文書とタームを同時にクラスタ化するクラスタリング手法を用い，名詞部分列をタームとし，適切な重みを設定し，ターム間の階層関係に基づき特徴表現を整理することで，個々のクラスタの特徴を反映した簡潔で分かり易い特徴表現付与を得た．

特徴表現付与に関して，今回は構成語が含まれる場合の上位下位関係だけを使って特徴表現を絞り込んだが，シソーラスのような単語間の上位下位関係を使っての絞り込みも同様にできると考える．今後の課題として，特徴表現間やクラスタ間の関係の分析などを行うことで，特徴表現をより絞り込むことが挙げられる．

参考文献

[1] Frakes, W. and Baeza-Yates(Editor), R.: *Information Retrieval : Data Structures & Algorithms*, Prentice Hall (1992).

[2] 湯浅夏樹, 上田徹, 外川文雄：大量のデータから自動抽出した名詞間共起関係による文書の自動分類, 情処研報 NL 98-10, pp. 81–88 (1993).

[3] 津田宏治, 仙田修司, 美濃導彦, 池田克夫：共起行列の固有ベクトルを用いる単語クラスタリング法 ～文書データベースの概要を表す単語クラスタの抽出 ～, 情処研報 NL 103-6, pp. 41–48 (1994).

[4] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R.: Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391–407 (1990).

[5] Golub, G. and Loan, C. V.: *Matrix Computations*, Johns-Hopkins, second edition (1989).

[6] 西野文人：日本語テキスト分類における特徴素抽出, 情処研報 NL 112-14, pp. 95–102 (1996).

- 解析, 形態素, 確率, モデル, 構造, 日本語
- 対話, 実, 発想, 発話, 目的, 管理, ユーザ, 支援, モデル
- 文, 検索, 法律, 構造, 要件, 効果, マルチメディア, システム, 理解, 自然言語

図 2: 名詞単語の場合の特徴表現付与例

- 解析, 形態素, 確率, モデル, 形態素 - 解析, 構造, 日本語
- 対話, 対話 - モデル, 実, 発話, 対話 - 管理, 目的, 支援, モデル
- 検索, 文, 法律, 構造, マルチメディア, システム, 類似

図 3: 名詞単語 + 名詞句の場合の特徴表現付与例

- 解析, 形態素 - 解析, 形態素, 日本語 - 形態素, 日本語 - 形態素 - 解析, 確率 - モデル, 確率, モデル, 日本語, 構造
- 対話, 対話 - システム, 対話 - モデル, 発話, 対話 - 管理, 実 - 対話, 目的, 理解, モデル, 自然言語
- 法律, 要件, 法律 - 文, 構造, 要件 - 効果 - 構造, 要件 - 効果, 効果 - 構造, 効果, 文, 自然言語

図 4: 名詞部分列の場合の特徴表現付与例

- 形態素 - 解析, (日本語 - 形態素 - 解析), 確率 - モデル, 構造

順位	類似度	文書 ID	論文タイトル
1	(0.67)	NL124-06	誤り駆動型の確率モデル学習による日本語形態素解析
2	(0.67)	NL123-01	枝分かれ構造をもつ連接確率モデルによる形態素解析
3	(0.67)	NL119-13	コスト最小法と確率モデルの統合による形態素解析

- 対話, (対話 - システム), (対話 - モデル), 発話, (対話 - 管理), (実 - 対話), 目的, 理解, 自然言語

1	(0.87)	NL126-18	WWWを介した対話システムとの対話における混乱の分析
2	(0.80)	NL114-19	反射と熟考の相互作用に基づく協調的対話モデル
3	(0.78)	NL126-17	情報の授受に基づく対話モデルについて

- 法律 - 文, 構造, (要件 - 効果 - 構造), 自然言語

1	(-0.64)	NL124-01	法律条文のデータ構造
2	(-0.64)	NL117-18	表層的手がかりによる六法全書法律文での要件部・効果部の抽出手法
3	(-0.63)	NL115-04	要件効果構造に基づく法律文制限言語モデルと法律文解析

- 情報 - 抽出, テンプレート, 形態素 - 解析

1	(0.58)	NL124-13	トップダウンなパターン解析に基づく情報抽出
2	(0.55)	NL125-03	数値情報をキーとした新聞記事からの情報抽出
3	(0.54)	NL115-12	テンプレートを用いた新聞記事からの製品情報抽出システム

- 分類, (自動 - 分類,) シソーラス

1	(-0.53)	NL117-14	シソーラスを用いた文書データの自動分類法
2	(-0.53)	NL120-09	係り受け関係を用いた副詞の分類と分類要素についての実験的評価
3	(-0.53)	NL123-09	分類パターンを用いた文書データの自動分類法

- 検索, (検索 - システム,) (文書 - 検索,) スコア, 類似, 的

1	(0.67)	NL127-04	適応型WWW自動検索手法
2	(0.59)	NL127-05	ユーザの情報利用目的に基づく検索システム
3	(0.53)	NL118-10	WWW環境での手話単語の検索システムの構築方法

- 辞書, 日英, 単語, 対訳, 翻訳, 機械翻訳, 処理

1	(-0.42)	NL122-03	共起関係を利用した対訳コーパスからの連語の対訳表現抽出
2	(-0.39)	NL128-10	日本語複合語の自動分割と日英語基対訳辞書の作成
3	(-0.35)	NL116-01	複数辞書の統合的利用のための汎用日本語辞書の構築

- ニュース, 字幕, (テレビ - ニュース - 番組,) 原稿, 要約, (自動 - 要約,) 分割, 記事, ＴＶ - ニュース

1	(-0.66)	NL122-13	テレビニュース番組の字幕に見られる要約の手法
2	(-0.63)	NL119-06	「テレビニュース番組電子化原稿を題材とした自動要約手法の大規模評価」
3	(-0.60)	NL126-09	短文分割を利用したテレビ字幕用自動要約

図 5: 特徴表現整理後の特徴表現付与例