7S-03

テキスト分類のための文書拡張法の評価

鳥山 修平 † 世木 博久 †
† 名古屋工業大学 情報工学科

1 はじめに

Web 上の twitter やブログなど短くスパースな文書を分類する際に, スパースな情報を補う方法として文書拡張 (あるいはターム拡張) 法が知られている [1, 2]. これらの研究は，形式概念分析の方法に基づくアプローチである．一方，[3] では，LDA(Latent Dirichlet Allocation)[4] に基づく文書拡張法を提案している．本稿では，この形式概念に基づく方法と LDA に基づく方法の二つの文書拡張法についてその分類結果を比較する．また，新たな単語間類似度を用いて, その文書拡張の効果を実験により確認する.

2 テキスト分類のための文書拡張

従来のテキストの分類やクラスタリングについては自然言語処理や情報検索の分野で多くの研究の蓄積がある．しかし，twitter やブログなど短くスパースな文書を扱う場合には, 単語の共起関係や共通する文脈情報がが乏しいため, 従来方法では必ずしも十分でない場合がある.

そこで，[1, 2] では形式概念分析 (FCA) の方法に基づいて, 単語文書行列 (term-document matrix, TDM) を形式概念における文脈と考えて, それから構築される単語概念束の特徴を利用した文書拡張法を提案している.

単語概念束による文書拡張

単語文書行列 TDM が与えられたとき, 単語 m が出現している文書の集合を m'，それらの文書集合に共通して出現している単語の集合を m'' と書く．このとき, 組 (m', m'') は単語概念であり, その全体は束をなす．図 1 にテキストから作られる TDM(文脈) と単語概念束の例を示す．すなわち, 文書の集合 $\{o_1, \ldots, o_4\}$ を形式概念におけるオブジェクト集合, 単語の集合 $\{a, \ldots, g\}$ を属性集合として扱う.

Boutari ら [2] は，単語 m_1, m_2 間の類似度 Sim_D を以下の式で与えている:

$$Sim_D(m_1, m_2) = \frac{1}{2} \cdot \left(\frac{2|m_1' \cap m_2'|}{|m_1'| + |m_2'|} + \frac{2|m_1'' \cap m_2''|}{|m_1''| + |m_2''|}\right).$$

Evaluating Text Expansion Methods for Short Text Classification

† Shuhei Toriyama (27115107@stn.nitech.ac.jp)
† Hirohisa Seki (seki@nitech.ac.jp)
Dept. of Computer Science, Nagoya Institute of Technology (†)
Showa-ku, Nagoya, 466-8555 Japan

	a	b	c	d	e	f	g
o_1	1	0	0	0	1	0	0
o_2	0	0	1	1	0	1	1
o_3	0	0	0	1	1	0	1
o_4	0	1	0	1	0	0	0

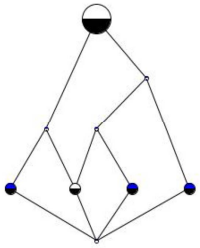

図 1: 例: テキストから作られる文脈 (左) と単語概念束 (右)

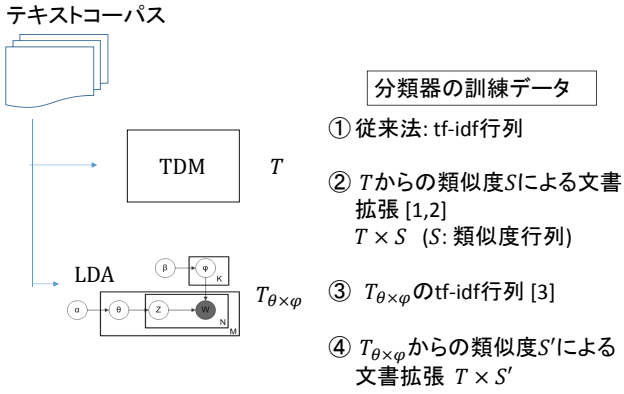

図 2: 本研究で用いる文書拡張方法

すなわち，$Sim_D(m_1, m_2)$ は，集合間の類似度として Dice 係数を考えたとき, オブジェクト (文書) 集合間の類似度 (第 1 項) と属性 (単語) 集合間の類似度 (第 2 項) の平均である.

LDA を用いた文書拡張

LDA[4] は自然言語処理のトピックモデルで使われる方法である．LDA におけるトピック分布 θ と単語分布 ϕ から作られる行列 $\theta \times \phi$ を考えると, それは拡張された単語文書行列 $T_{\theta\times\phi}$ と考えることができる．[3] では，この $T_{\theta\times\phi}$ を分類器の訓練データとして用いる方法を提案している.

3 単語間類似度と評価方法

本研究で用いる単語間類似度

本研究では, 単語間類似度として $Sim_D(m_1, m_2)$ における Dice 係数の代わりに，もう一つの代表的な集合間類似度である Jaccard 係数を用いた類似度尺度 $Sim_J(m_1, m_2)$ を用いて, それを利用した場合の文書拡張効果を調べる.

表 1: 実験対象のコーパス

(1) Reuters-21578 R8 の部分データセット

Class	#(train)	#(test)	Total
acq	1596	696	2292
crude	253	121	374
earn	2840	1083	3923
Total	4689	1900	6589

(2) Google Snippets コーパスの部分データセット

Class	#(train)	#(test)	Total
business	1200	300	1500
culture-arts-ent.	1880	330	2210
ed.-science	2360	300	2660
Total	5440	930	6370

更に，LDA によって得られる拡張された TDM: $T_{\theta\times\phi}$=$\theta\times\phi$ から，単語間類似度を重み付き Jaccard 係数を用いて次のように定義する: $Sim_{wJ}(m_1, m_2) = \Sigma_k \min(m'_1, m'_2)/\Sigma_k \max(m'_1, m'_2)$. ここで，$m'_1$ は行列 $\theta\times\phi$ における m_1 に対する列ベクトル $(w_1, \ldots, w_n)$ (n: 全文書数) で, w_i は文書 i に単語 m_1 が出現する確率である．m'_2 も同様．本研究では，この Sim_{wJ} を単語間類似度として文書拡張を行う．

評価方法

実験に用いるデータセットは, 先行研究 [2, 3] でも使われているコーパスを対象とした (表 1). 一つ目は Reuters-21578 の中の単一ラベルをもつ R8 データセット [5] から，数の多い上位 3 クラスの文書集合を用いた．本研究ではごく短いテキストの扱いに関心があるので, 各ニュースの見出し (title field) だけを入力文書とした．二つ目は，Google Snippets コーパスから抽出した表 1 に示す部分データセット†を用いた．

テキスト分類のための分類器としては k-NN(scikit-learn) を用いた．また, ハイパーパラメータの近傍点数 k とトピック数 N については，実験により $k = 19$, $N = 10$ とした．

4 実験結果

表 2 に各方法によるテキスト分類精度の実験結果を示す．表で，*tf-idf* は従来の tf-idf 法による分類, Sim_D (Sim_J) はそれぞれ Dice(Jaccard) 係数による文書拡張を行った分類, $tf\text{-}idf_{LDA}$ は $T_{\theta\times\phi}$ に対する tf-idf 法による分類， Sim_{wJ} は $T_{\theta\times\phi}$ から重み付き Jaccard 係数で文書拡張した分類法を示す．表の分類精度は重み付き平均 (weighted average) の値である．

† http://jwebpro.sourceforge.net/data-web-snippets.tar.gz

表 2: 実験結果

(1) Reuters-21578 R8 の部分データセット

	tf-idf	Sim_D	Sim_J	$tf\text{-}idf_{LDA}$	Sim_{wJ}
適合率	0.860	**0.862**	**0.862**	0.787	0.833
再現率	0.845	**0.855**	0.852	0.802	0.808
F-尺度	0.839	**0.852**	0.848	0.782	0.784

(2) Google Snippets コーパスの部分データセット

	tf-idf	Sim_D	Sim_J	$tf\text{-}idf_{LDA}$	Sim_{wJ}
適合率	0.834	**0.847**	0.840	0.827	0.838
再現率	0.824	**0.840**	0.832	0.831	0.822
F-尺度	0.824	**0.840**	0.832	0.825	0.801

実験では, いずれの場合も基準となる従来法 (*tf-idf*) と比較して二つの文書拡張法 Sim_D, Sim_J が良い結果を与えている．それに対して, LDA の拡張行列を用いた方法 ($tf\text{-}idf_{LDA}$) は十分な効果が得られず，データセット (1) では従来法よりも悪くなっている．

5 まとめ

本研究では, 短くスパースな文書の分類のための文書拡張法における形式概念分析の方法と LDA による文書拡張法に基づき，単語間類似度 Sim_J, Sim_{wJ} を導入し文書拡張の効果を実験した．LDA を用いた文書拡張法は他の方法に比べその有効性が十分でないという結果であった．今後の課題として, 他のコーパスや，異なる分類器・クラスタリング法による実験とその評価が挙げられる．

謝辞 本研究は JSPS 科研費 (C)18K11432 を受けたものです．

参考文献

[1] Carpineto, C. et al.: A Concept Lattice-Based Kernel for SVM Text Classification, *FCA*, Springer Berlin Heidelberg, pp. 237–250 (2009).

[2] Boutari, A. M. et al.: Evaluating Term Concept Association Mesaures for Short Text Expansion: Two Case Studies of Classification and Clustering, *CLA* (2010).

[3] Rogers, N. et al.: A Comparison on the Classification of Short-text Documents Using Latent Dirichlet Allocation and Formal Concept Analysis, *AICS 2017*, pp. 50–62 (2017).

[4] Blei, D. M. et al.: Latent Dirichlet Allocation, *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022 (2003).

[5] Cardoso-Cachopo, A.: Improving Methods for Single-label Text Categorization, PdD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa (2007).