九州大学学術情報リポジトリ
Kyushu University Institutional Repository

単語の分散表現を用いた文書群のラベル推定

加登，一成
九州大学工学部電気情報工学科

伊東，栄典
九州大学情報基盤研究開発センター

https://hdl.handle.net/2324/2544135

出版情報：電気・情報関係学会九州支部連合大会講演論文集. 72 (10-2P-04), pp.580-581, 2019-09-28.
電気・情報関係学会九州支部
バージョン：
権利関係：©2019 電気・情報関係学会九州支部連合大会委員会

単語の分散表現を用いた文書群のラベル推定

加登一成* 伊東栄典**
（九州大学 *工学部電気情報工学科 **情報基盤研究開発センター）

1 はじめに

文書群をクラスタリングで同類の部分文書集合に分割する際，出力後の部分文書集合の種類や意味は，人間が読んで判定していた。この判定を機械的に行いたい。本研究では，SVM 分類器の重要語と，単語の分散表現が近い単語をラベル候補する手法を提案する。

実験として，分類問題用のラベル付き文書集合を用いて，部分文書集合からラベルを推定した。具体例にライブドアニュースコーパスを用いた。

2 ライブドアニュースコーパス

ライブドアニュースコーパス[1] は、NHN Japan 株式会社が運営する livedoor ニュースを収集したものである。文書は表1に示す 9 つのカテゴリに分かれている。各文書は URL、作成日時、タイトル、本文からなる構成である。

表1 ライブドアニュースコーパスの文書内訳

カテゴリ	文書数
独女通信	870
Sports Watch	900
家電チャンネル	864
MOVIE ENTER	870
トピックニュース	770
IT ライフハック	870
エスマックス	870
livedoor HOMME	511
Peachy	842

3 fastText

fastText[2] は Facebook AI Research が 2016 年に開発した自然言語処理向けアルゴリズムであり、GitHub にてオープンソースとして公開している単語のベクトル化とテキスト分類をサポートした機械学習ライブラリである [3]。単語の分散表現を獲得し高次元のベクトルで表現する。分散表現では、vector(‘king’) – vector(‘man’) + vector(‘woman’)が vector(‘queen’)に近似するような加法・減法が成り立つ規則性が示されている [4]。本研究では、fastText が使用するモデルの内、文章中に含まれる単語の並びから単語の出現確率を利用する Skip-gram モデルを用いて分散表現を獲得する。

4 ラベル推定の方法

本研究では、ラベル推定問題を2つの部分問題に分割する。1つ目は、SVM での文書クラスの重要語抽出である。2つ目は、重要語からのラベル語推定である。

4．1 SVM を用いた重要語の導出

SVM は 1995 年頃に AT&T の V.Vapnik が発表したパターン識別用の教師あり機械学習方法である。マージン最大化で汎化能力が高く、分類器の中でも高性能かつ高速な識別を可能にする。データの2クラス分類に秀でており、多クラス分類も2クラス分類を複数回行うことで対応できる。

N 個の文書からなる文書集合 D を考える。文書 $d\ (d \in D)$ が属するクラスも与える。各文書に登場する単語を抽出し、文書 d を Bag of Words で表現する。更に各単語が文書 d に登場するか否かを調べ、d を単語ベクトルとして表現する。最終的に、全文書を文書単語行列で表す。

次に、文書単語行列を学習データとして線形 SVM を用いて文書分類器を作成する。文書分類器は文書 d があるクラス C に属すか否かを判定する。文書分類器により、クラス C に対する単語の重みを得る。正の重みを持つ単語は正例に影響が大きく、その絶対値が大きい程クラス C と関連が深いと言える。ここでは、正の重みが大きい単語上位 K 個を重要語とする。

4．2 重要語からのラベル候補選出

クラス C の重要語の集合を V とする。クラス C のベクトルは重要語から算出する。具体的には、重要語の単語ベクトルの平均を取る。計算を以下に示す。

$$vec(C) = \frac{1}{|V|}\sum_{x \in V} vec(x)$$

最後に、クラス C のベクトル $vec(C)$ から、クラス C のラベル候補を選出する。日本語 wikipedia の記事の名詞だけの文章を fastText で学習させて得られた単語ベクトルを用いて、単語 x のベクトル $vec(x)$ とクラス C のベクトル $vec(C)$ のコサイン類似度を計算する。計算は以下である。

$$\cos(C, x) = \frac{vec(C) * vec(x)}{|vec(C)||vec(x)|}$$

コサイン類似度が高い順に単語を並べ、上位の単語をクラス C のラベル候補とする。

5 実験

本研究では、文書に含まれる抽出対象の単語を名詞のみにし、全文書中に3文書以上かつ全文書の半分の文書以下に登場する単語に限定した。また、ニュースカテゴリそれぞれをクラスとし、独女通信・Sports Watch・家電チャンネル・MOVIE ENTER の4クラス分の文書分類器を作った。

SVM により算出された各クラスの重要語上位10単語と $K = 10$ としたときの各クラスのラベル候補を表2と表3に示す。但し、独女通信の重要語である「独女」・「オフィスエムツー」は単語ベクトルが存在しなかったため、11・12位の「10周年」・「記念に」をラベル推定に使用した。

6 考察

まず、SVM での重要語抽出において考える。単語数が極端に少ない文書があり、その文書に登場する単語が重要語として現れている。また、文書の著者等、あるクラスの文書では頻出であるにも関わらずラベル推定に寄与しないと思われる単語が重要語なので、SVM に学習させる前に不要な部分や文書を切り捨てる方が良いと考える。

次に、ラベル推定において考える。独女通信は話題が広いためか、ラベル候補も広いジャンルのものになると考える。他の3クラスでは、それぞれのクラスに関連する固有名詞が多い。これはクラス内の文書の話題が特定分野の話題に限定されているからであろう。

7 おわりに

本研究では単語の分散表現を用いて文書群のラベル推定を行った。wikipedia に含まれる名詞という膨大な候補の中からラベル候補を見つけることで各クラスの上位概念を探そうとしたが、得られた結果は上位概念とはいえないものだった。今後の課題として、単語ベクトルを学習するための文書集合を別の物に変えることを検討している。

参考文献

[1] RONDHUIT, ダウンロード, http://www.rondhuit.com/download.html#ldcc, 参照 Jul.26,2019.

[2] facebookresearch, fastText, https://github.com/facebookresearch/fastText, 参照 Jul.26,2019.

[3] NISSEN DIGITAL HUB, Facebook が開発した fastText とは？その活用事例を解説, https://nissenad-digitalhub.com/articles/facebook-fasttext/, 参照 Jul.26,2019.

[4] T. Mikolov, I. Sutskever, K. Chen, G. Corado, and J. Dean : Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.

表2 各クラスの重要語

	独女通信	Sports Watch	家電チャンネル	MOVIE ENTER
1	独女	Sports	話題	映画
2	オフィスエムツー	Watch	本日	征服
3	オトナ女子	インターネット上	売れ筋	スカイライン
4	境界線	選手	関連	DVD
5	Style	ファン	ネット	本作
6	BIGLOBE	戦	家電	MOVIE
7	平気	ロンドン五輪	パナソニック	ENTER
8	6 月 9 日	美女	亜紀子	特集
9	MIWA	氏	牧田	公開
10	HARD	サッカーファン	1	和製

表3 各クラスのラベル候補

	独女通信	Sports Watch	家電チャンネル	MOVIE ENTER
1	EXHiBiTiON	US サッカーアスリートオブザイヤー	スマイる No.1 ショップ	MOVIE 輝きの向こう側へ!
2	DEYEGIRL	マッチデーハイライト	パナソニックセールスマンカタログ	オリジナルムービー
3	SHOWNEN	オフィシャルツイッター	テレビシャカイ実験あすなろラボ	MOVIE-
4	ChageLiveTour	サッカーファン	家電	銀魂 2 掟は破るためにこそある
5	THE ウラ BEST!私だけのドリカム	SHERDOG	カスタムインイヤモニター	MOVIES
6	HIXNADE	Goal.com	住商ホームショッピング	Hi☆sCoool!セハガール
7	高見沢俊彦のロックばん	FIFPro	パナソニックショップ	劇場版弱虫ペダル
8	たまゆら~もあぐれっしぶ~	SportFight	新型テレビ	スピンオフネットムービー
9	しおりごと-BEST-	NBC スポーツ	パソコンサンデー	劇場版仮面ライダーゴースト 100 の眼魂とゴースト運命の瞬間
10	SHOWGATE	SportsCenter	ハイエンドテレビ	劇場版七つの大罪天空の囚われ人