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

大規模文書コーパスから得た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/2324/2557145

出版情報：情報処理学会全国大会講演論文集. 82, pp.7T-03-, 2020-03-06. 情報処理学会
バージョン：
権利関係：ここに掲載した著作物の利用に関する注意：本著作物の著作権は情報処理学会に帰属します。本著作物は著作権者である情報処理学会の許可のもとに掲載するものです。ご利用に当たっては「著作権法」ならびに「情報処理学会倫理綱領」に従うことをお願いいたします。

大規模文書コーパスから得た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†　　九州大学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，人間が文書を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。具体例として，日本語ラベル付き文書集合のライブドアニュースコーパスを用いた。本論文では提案手法と用いた文書集合を説明し，実験結果の考察についても述べる。

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る。GitHub にてオープンソースとして公開されており、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 の重要語を上位から $t_1, t_2, \dots, t_K$ とする。次に、日本語 Wikipedia の記事の名詞だけを fastText で学習させて単語ベクトルを獲得し、この単語集合を X とする。t_1 と X に含まれる全単語についてコサイン類似度を計算し、類似度上位の単語 n 個を求める。計算は以下である。

$$\cos(t_1, x) = \frac{vec(t_1) * vec(x)}{|vec(t_1)||vec(x)|}, (x \in X)$$

t_2からt_Kも同様に、それぞれn個の単語を求める。重要語tの SVM での重みをw_tとしたとき、単語xのスコアを以下のように定義する。

$$score(x) = w_t * \cos(t, x)$$

重要語K個それぞれに対しコサイン類似度上位の単語n個を求めたので、得られた集合は単語数$K * n$個になる。この集合から前述のスコア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た。

SVM により算出された各クラスの重要語上位 10 単語を表２に，$K = 10, n = 10$ としたときの各クラスのラベル候補を表３に示す。

6. 考察

まず、SVM での重要語抽出において考える。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。it ライフハックは上位の単語が Togetter とそれに似た単語になっており、ラベル候補としてふさわしいとはいえない。他の３クラスでは、上位にクラス名に近い単語が現れている。it ライフハックに関しては文書中の話題が他クラスよりやや広いためであろう。

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。しかし、全てのクラスについて妥当なラベルが得られたとはいえなかった。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

	sports watch	家電チャンネル	movie enter	it ライフハック
1	Sports	話題	映画	Togetter
2	Watch	本日	征服	クチコミ
3	インターネット上	売れ筋	スカイライン	筆者
4	選手	関連	DVD	2012 年
5	ファン	ネット	本作	販売元
6	戦	家電	MOVIE	hack
7	ロンドン五輪	パナソニック	ENTER	life
8	美女	亜紀子	特集	IT
9	氏	牧田	公開	昨日
10	サッカーファン	1	和製	モノ

表３　各クラスのラベル候補

	sports watch	家電チャンネル	movie enter	it ライフハック
1	Sports	話題	映画	Togetter
2	Watch	本日	征服	getter
3	eSports	売れ筋	スカイライン	Together
4	Sportswear	関連	映画作品	√letter
5	Sportsmen	注目	DVD	Getter
6	Sportscar	話題性	動物映画	クチコミ
7	Sporty	マスコミ	マサラ映画	better
8	Sportsman	静かなブーム	学園映画	Wetter
9	SportsCenter	ネットで話題	バカ映画	letter
10	Sport	今週、妻が浮気します	北野映画	setter