単語の分散表現を用いた文書クラスタのラベル推定

Document cluster label estimation using word vectors

淀川 翼[1]　加登 一成[2]　伊東 栄典[3]

Tsubasa Yodogawa[1], Issei Kato[2], and Eisuke Ito[3]

[1]九州大学大学院ライブラリーサイエンス専攻

[1] Graduate school of Library Science, Kyushu University.

[2] 九州大学工学部電気情報工学科 School of Engineering, Kyushu University

[3] 九州大学情報基盤研究開発センター Research Institute for IT, Kyushu University

Abstract: Clustering is applied to divide customers into small subsets to analyze detail of customers. The attributes of each subsets are manually researched by human analysts. In case of documents, human analysts can extract the attributes of each subdocument set generated by clustering, if they check all documents. However, it is not practical to analyze by human if the size of documents is large. So, mechanical attributes estimation is required. In this paper, we propose a method for estimating the labels. The proposed method consists of three steps. At first, obtain distributed representation of words by fastText and large corpus. Next, extract feature words using SVM discriminator. Finally, estimate appropriate labels of a document set using similarity of word vector and SVM feature words. As an experiment, we apply proposed method to two document sets, The 20 newsgroups and Livedoor news corpus. Both sets are used for classification problem. We report the proposed method and experimental results.

1. はじめに

教師なし機械学習手法であるクラスタリング手法として階層化手法の Ward 法や，非階層化手法の K−means 法などが検討されてきた．クラスタリングの問題として，生成されたクラスタの解釈がある．従来は計算結果として生成されたクラスタを人の手で分析し，そのクラスタの意味を持たせていた．たとえば顧客分析では，クラスタリングされた顧客の集合に分析者が意味を与え，それを企業でのニーズ分析や売り上げ向上などのマーケティングに役立てていた．また，楽曲・映像・漫画小説などのコンテンツ集合のクラスタリングでは，コンテンツ集合の意味を与えることでコンテンツ推薦に役立てることが出来る．

クラスタリングを適用した際，クラスタは何らかの属性を持つと考えられる．しかし，分割されたクラスタの属性を把握する手法は確立されていない．分割されたクラスタに対して，クラスタ属性を示すラベルを機械的に付与できれば，人手によるクラスタ属性の確認作業無しに，クラスタが目的に沿って正しくクラスタリングされているかを評価できる．

既存の研究としてクラスタリングされたニコニコ動画の動画メタデータ文書群へのラベル付けや[1], Naive Bayes モデルを用いた手法[2]などがあるが, 本研究ではその前段階として，正解ラベル付き文書データ群に対し，文書の内容からのラベル推定を行う．正解ラベル付き文書データとして 1995 年に Ken Lang により公開された The 20 Newsgroups data set と，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ライブドアニュースコーパスを用いた.本研究では文書集合のラベル推定のために，Facebook AI が提案・公開している単語の分散表現手法である fastText を用いた[3]. Facebook AI は手法とツールを公開するだけでなく，英語版 Wikipedia を文書コーパスとする単語ベクトル(wiki.en.vec)を公開している[4]．The 20 Newsgroups data set を用いた実験ではこの英語版 Wikipedia から生成された単語ベクトルを用いた．また，ライブドアニュースコーパスを用いた実験では, 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した．

本研究では，文書集合のラベル推定方法として単語分散表現を用いた手法を提案する．SVM を用いた文書分類で算出された単語の重みを用いて文書群の

重要語を定義し，その重要語と関連する語を文書群のラベルとする方法である．

2. 用いたデータ

本研究では文書群に対し，文書の内容からのラベル推定を行う．そのため予め正解ラベルが付与された文書群を用いる．分析対象とする正解ラベル付き文書集合として， The 20 Newsgroups データセットとライブドアニュースコーパスを用いる．

2.1 The 20 Newsgroups

The 20 Newsgroups データセット[5]は，1995 年に Ken Lang により公開されたものである．表 1 に示す 20 個のニュースグループに投稿された Usenet のニュース記事を集めたものである．Usenet とは，インターネット上に提供された分野別のニュース記事投稿提示サービスである．表 1 に文書数を示す．

表 1. The 20 Newsgroups の文書数

グループ	数	グループ	数
alt.atheism	480	rec.sport.hockey	598
comp.graphics	581	sci.crypt	594
comp.os.ms-windows.misc	572	sci.electronics	591
comp.sys.ibm.pc.hardware	587	sci.med	594
comp.sys.mac.hardware	575	sci.space	593
comp.windows.x	592	soc.religion.christian	599
misc.forsale	582	talk.politics.guns	545
rec.autos	592	talk.politics.mideast	564
rec.motorcycles	596	talk.politics.misc	464
rec.sport.baseball	594	talk.religion.misc	376

2.2 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[6]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. 表2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．

3. ラベル推定手法

ラベル推定問題を２つの部分問題に分割する． 1 つ目は文書クラスの重要語抽出問題である． ２つ目は, クラスの重要語からのラベル語推定問題である. 手法2では, SVM で文書クラスの重要語抽出を行い, その後に重要語からラベル語を推定する．

表 2. ライブドアニュースコーパスの文書数

カテゴリ	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

3.1 SVM を用いた重要語抽出

SVM (Support Vector Machine) を用いた文書クラスの重要語抽出について説明する．

3.1.1 SVM (Support Vector Machine)

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り，局所解収束に関する問題が無い．マージン最大化で汎化能力を高めており，現在知られている分類器として高速かつ高性能な識別能力を持つ．線形でない非線形カーネルも利用可能であるため，線形分離不可能な分類問題にも適用可能で応用範囲が広い．データを 2 つに分類する２クラス分類には優れている．多クラス分類は，２クラス分類を複数回適用することで対応できる．

3.1.2 線形 SVM による 2 クラス文書分類

重要語抽出のために用いた線形 SVM による２クラス文書分類を説明する．N 個の文書から成る文書集合 D がある．文書 d ($d \in D$)が属するクラスも与えられる．各文書 d の中に出現する単語を抽出し，文書 d を Bag of Words で表現する．更に各単語の出現頻度を数え上げることで，d を単語の頻度ベクトルで表現できる．これにより全文書を文書単語行列 (document word matrix)で表現する．ここまでの手順を図 1 に示す．

次に，文書単語行列を学習データに用いて線形 SVM の文書分類器を作る．文書分類器は，あるクラス C に属する文書か否かを判定する．本研究では２クラス分類の線形 SVM を用いるため，文書クラス数と同数の SVM 分類器を作成する．学習データである文書単語行列を SVM で学習させることで，クラス C に対する単語への重みが算出される．

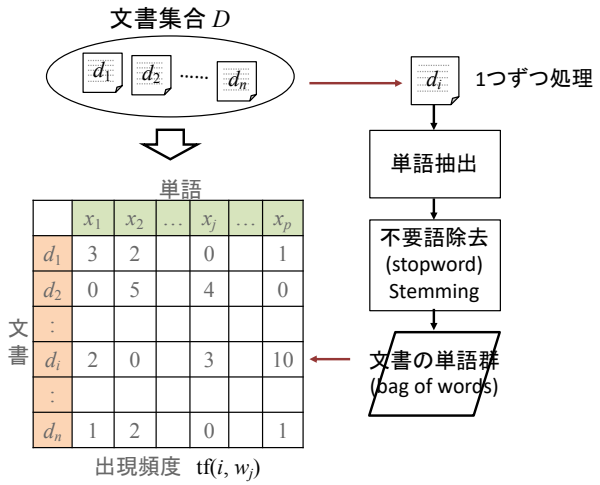

図 1 文書集合と Bag of Words，文書単語行列

3.1.3 文書クラスに対する重要語選出

クラス C の SVM 文書分類器作成により，クラス C に関する単語の重みが算出される．正の重みを持つ単語は正例に影響が大きいため，クラス C と関連が大きい．逆に負の重みを持つ単語は負例に影響が大きいため，クラス C と関連が小さい．重みの絶対値が大きいほど影響が大きい．そこで，正の重みの大きな単語をクラス C に対する重要語とする．本実験において重要語は正の重み上位の K = 10 個とした．

3.2 重要語からのラベル候補選出

次にクラス C に対する重要語から，本研究の目的であるクラスのラベル推定手法を述べる．

SVM が算出した単語の重みを用いて，重みが上位 K 個の単語をクラス C に対する重要語とする．この重要語集合を Vc とする．クラス C の重要語を用いてクラス C のベクトルを算出する．クラス C のベクトルは，K 個の重要語の単語ベクトルの平均値とする．各単語のベクトルは，コーパスからの学習で得た単語ベクトルを用いる．式にクラス C のベクトル算出を示す．

$$vec(C) = \frac{1}{K}\sum_{x \in Vc} vec(x)$$

最後に，算出したクラス C のベクトル vec(C) から，文書クラス C のラベル候補を選出する．コーパスからの学習で得た単語ベクトルを用いて，クラス C のベクトル vec(C)と単語 x のベクトル vec(x)との類似度を計算する．類似度としてはコサイン類似度を用いる．計算は以下である．

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

類似度の大きな順に単語を並べ，類似度の上位個を文書クラス C のラベル候補とする． 図 2 に SVM での単語の重み算出と分類器作成とを示す．

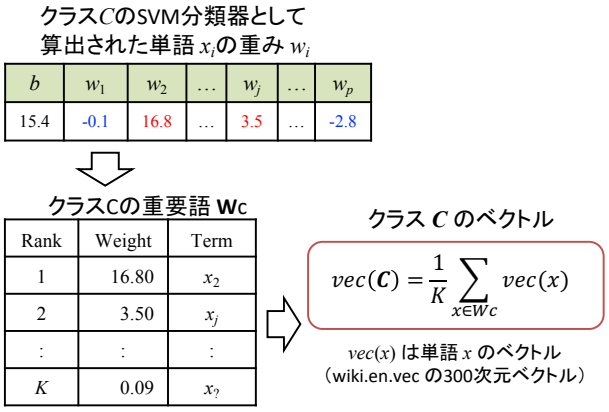

図 2 クラス C の重要語からのベクトル算出

4. 実験

4.1 The 20 Newsgroups

The 20 Newsgroups の文書集合に，推定手法を適用した．SVM により算出された各クラスの重要語上位 10 単語と，クラスごとの上位 10 単語の一部を示す．紙面では都合上 4 グループに絞って示す．ただし，各クラスの重要語に含まれる単語のうち Wiki.en.vec に含まれないものは上位語から除外して 11 番目以降の単語を繰り上げて示している．

4.2 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を用いた研究では，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．こちらも SVM により算出された各クラスの重要語上位 10 単語と各クラスのラベル候補を表に示す．ただし，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，こちらも 11 番目以降の単語を繰り上げて示している．

5. 考察

The 20 newsgroups を用いた実験では，ニュースグループ名と関連する単語が多く出現した．例えば

alt.atheism グループでは，ラベル候補に近い単語が多く上位 10 単語に存在している．また misc.forsale グループでは重要語に存在しなかった buy や purchase といった単語が上位に選出されている．このことは，SVM で抽出された語句内にふさわしい語が抽出されていなかったとしても，ラベル推定にふさわしい語句が出現する可能性があるといえる．

一方，comp.windows.x グループのラベル候補には，重要語に存在した windows が存在しなくなり，windows を含む周辺語が多く選出されている．また rec.sport.baseball グループも，重要語最上位であった baseball がなくなり，MLB のチーム名など固有名詞が上位に選出された．これらのことから，ラベル候補はラベルの語句の抽象度に左右されるということが推測できる．抽象度の低いものほど，上位概念よりも，固有名詞を含む下位概念を多く抽出する傾向にある．

ライブドアニュースコーパスを用いた実験では，独女通信で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った．一方他の 3 クラスでは，The 20 newsgroups と同様にそれぞれのクラスに関連する固有名詞が多くなっている．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ためであるためと考えられる．

6. 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，得られた結果の多くは上位概念とはいえないものであった．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考えている．また，SVM が導出した重みの上位 K の値を変化させた場合や負の重みを持った語句を計算に含めて比較を行うなど，適切なパラメータの導出の分析も行いたい．そして，ラベル候補がどれほどふさわしいか定量的な評価の確立を行うことを検討している．

参考文献

[1] 飯田委哉，伊東栄典，佐嘉田悠樹:クラスタリングによるオンライン小説の多様性動向分析,火の国情報シンポジウム論文集, pp.1-7 (2018).

[2] 小島諒介，亀谷由隆，佐藤泰介:Naive Bayes モデルを用いた効率的なクラスタラベリング手法，人工知能学会人工知能基本問題研究会資料(SIG-FPAI-B), Vol.88 pp. 19-24, (2013)

[3] Bojanowski P., Grave E., Joulin A., Mikolov T.: Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics, Vol.5, pp. 135-146, (2016)

[4] GitHub–facebookresearch/fastText, https://github.com/facebookresearch/fastText, (accessed at Nov.06, 2019)

[5] son Rennie: Home Page for 20 Newsgroups Data Set, http://qwone.com/~jason/20Newsgroups/ (accessed at Nov.06, 2019)

[6] RONDHUIT: ダウンロード, http://www.rondhuit.com/download.html, (accessed at Nov.06, 2019)

表 3. SVM による重要語上位 10 単語（The 20 Newsgroup）

	alt.atheism	comp.windows.x	misc.forsale	rec.sport.baseball
1	keith	motif	sale	baseball
2	benedict	xterm	offer	phillies
3	mathew	widget	shipping	sox
4	atheists	server	sell	cubs
5	atheism	xlib	obo	career
6	gregg	window	pay	pitcher
7	atheist	widgets	summer	mattingly
8	believing	openwindows	asking	ball
9	islamic	clients	offers	stadium
10	tammy	consortium	camera	mets

表 4. ラベル候補（The 20 Newsgroups）

	alt.atheism	comp.windows.x	misc.forsale	rec.sport.baseball
1	atheist	openwindows	buy	phillies
2	atheism	widgets	offer	yankees
3	atheisty	sqlwindows	purchase	astros
4	atheistrabbi	wxwidgets	sell	shortstop
5	atheists	xpwindows	sale	baseman
6	atheistic	wxwindows	purchases	mets
7	atheistical	qdesktopwidget	buying	sox
8	apatheist	decwindows	purchasing	outfielder
9	theist	openwindow	reselling	diamondbacks
10	atheistically	windowing	pay	dodgers

表 5. SVM による重要語上位 10 単語 (ライブドアニュースコーパス)

	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 6. ラベル候補（ライブドアニュースコーパス）

	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