CGM百科辞典を用いた利用者投稿動画クラスタリング

佐嘉田, 悠樹
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

廣川, 佐千男
九州大学情報基盤研究開発センター

出版情報：電気関係学会九州支部連合大会講演論文集, pp.544-545, 2017-09-28. 電気・情報関係学会九州支部
バージョン：
権利関係：許諾番号：17SB0111

CGM 百科辞典を用いた利用者投稿動画クラスタリング

佐嘉田悠樹 * 伊東栄典 ** 廣川佐千男 ***
(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)
* y.sakata.222@s.kyushu-u.ac.jp, ** ito.eisuke.523@m.kyushu-u.ac.jp, *** hirokawa@cc.kyushu-u.ac.jp

1 はじめに

近年，利用者がコンテンツを投稿するサービス（CGM, Consumer Generated Media）が人気である。YouTube やニコニコ動画などの動画 CGM サイトには多数の動画が投稿されるため，閲覧したい動画の選択が困難である。閲覧動画の推薦にはクラスタリグが有用である。本論文ではニコニコ動画について，動画のクラスタリングを実行した。各動画に付随するメタデータとして，動画投稿者が付与するタイトルと説明文と，投稿者と視聴者が付与できるタグが存在する。これらは重要であるものの情報量が少なく，また単語のゆらぎも有る。そこでインターネット百科事典の一つであるニコニコ大百科を利用した動画クラスタリングを提案する。ニコニコ大百科のページを取得し，Word2Vec を用いて単語のベクトル化を行い，実際のメタデータに適用し，K-means によるクラスタリング後，評価を行なう。

2 CGM 百科辞典を用いた単語のベクトル表現

本研究で用いた手法とデータについて説明する。図 1 にデータ処理の流れを示す。

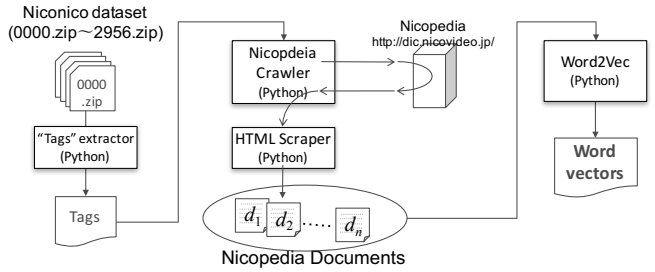

図 1: データ処理の流れ

2.1 Word2Vec，Doc2Vec

Word2Vec は Google の研究者 Tomas Mikolov のチームの開発した分散表現を生成するツールで，各単語を高次元のベクトルで表現する。単語をベクトル表現することで，例えば 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 [1]。Word2Vec は文章中に含まれる単語の出現数を利用する Bag-of-Words モデルと，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生成する。また，同様の手法を使用したものに Doc2Vec が存在する。こちらは，単語ではなく文書の分散表現を生成できる。

2.2 ニコニコ大百科を用いた Word2Vec による単語ベクトル化

Word2Vec や Doc2Vec を用いる場合，単語を適切なベクトルで表現するための学習データが必要である。ニコニコ動画のメタデータに含まれる文章の分析には，ニコニコ動画に適した学習データが望ましいため，ニコニコ大百科を用いる。ニコニコ大百科は，株式会社大百科ニュース社（2017 年 7 月 3 日までは有限会社未来検索ブラジル）が運営するインターネット百科事典の一つである。そのため株式会社ドワンゴの運営するニコニコ動画と関わりがある内容を含む。なお，記事作成・編集はニコニコ動画プレミアム会員に限定されている。

ニコニコ大百科データを収集するプログラムを Python 言語を用いて作成した。多数存在するニコニコ大百科の記事から，動画タグ（単語）の説明ページ 239,753 件を収集した。HTML で記述された説明ページから，単語を説明する文章部分をプログラムで切り取り，それらを Word2Vec に適用する学習データ（コーパス）とした。Python 用の自然言語処理および機械学習モジュール群 gensim に含まれる Word2Vec[2] を使い，学習用データから単語の分散表現 (100 次元ベクトル) を生成した。

3 クラスタリング

クラスタリングでは，対象をベクトルで表現する必要がある。本研究では動画メタデータ（文書）群がクラスタリング対象である。動画メタデータのベクトルは，次式で算出した。各文書に出現する単語について，予め Word2Vec で算出した単語ベクトル値を取る。それを足し合わせ，文書の単語数で割る。この値を文書のベクトルとした。

$$vector(d) = \frac{\sum_{w \in d} vector(w)}{n}, \quad n \text{ は文書 } d \text{ の単語数.}$$

上式で得た文書ベクトルを K-means 法を用いてクラスタリングした。クラスタリング後の結果を t-SNE を用いて次元圧縮して結果を描画した。t-SNE 法 (t-Distributed Stochastic Neighbor Embedding : t 分布型確率的近傍埋め込み法) は Laurens van der Maaten と Geoffrey Hinton の開発した次元圧縮アルゴリズム [3] で，高次元のデータの可視化（低次元表現）に効果的である。K-means と t-SNE は，Python 用機械学習モジュール群 scikit-learn [4] に含まれるものを用いた。

クラスタリングの評価には，重なりを表すエントロピー，正解クラスタを含む割合を表す純度，再現率と精度の重み付き平均を表す F 尺度を用いた。いずれも 0 から 1 の値を取り，エントロピーは小さいほど，純度と F 尺度は大きい結果が良好であることを示す。

4 実験と考察

4.1 実験用データ

国立情報学研究所は，ドワンゴ社と未来検索ブラジルから社と協力し，ニコニコ動画の動画メタデータを提供している。全データセットには約 1400 万件の動画メタデータが有る。全データのクラスタリングは大規模で評価困難であるため，小規模の実験用テストデータを作成した。

表 1 に示すタグを含む動画から，再生数が 10000 回以上，かつタグの個数が 10 個であるものから，無作為に 100 個，合計 600 個の動画メタデータを選出した。視聴者が少ない動画は視聴者が付与するタグの質と個数の差が出ないように選出した。

表 1: 実験用データのタグ

FPS，Minecraft，政治
音楽，歌ってみた，踊ってみた

4.2 対象属性

動画メタデータは複数の属性を持つ。そこで (i) タグのみ，(ii) タイトルとタグ，(iii) 説明文とタグ，(iv) タイトル

と説明文とタグ全て，の 4 通りを検討する。

4.3 結果

(i)-(iv) の 6-means のクラスタリング結果をエントロピー，純度，F 尺度で評価した。(表 2) また各々を t-SNE で描画した。(図 1〜図 8：2-means〜5-means は省略) また結果が最良である (ii) の 6-means のクラスタリング結果を表 3 に示す。

表 2: クラスタリング評価

	(i)tags	(ii)title	(iii)dscrpt	(iv)all
Entropy	0.231	0.292	0.692	0.521
Purity	0.630	0.698	0.358	0.477
F-measure	0.641	0.700	0.345	0.467

表 3: (ii) の 6-means クラスタリング

	C1	C2	C3	C4	C5	C6
FPS	76	22	0	0	2	0
Minecraft	7	93	0	0	0	0
政治	0	0	96	0	4	0
音楽	1	0	0	45	41	13
歌ってみた	0	0	1	60	29	10
踊ってみた	2	0	0	4	57	37

5 考察

エントロピーでは (i) タグのみが，純度，F 尺度では (ii) タイトルとタグが良好な結果を示している一方，エントロピー，純度，F 尺度いずれでも (iii) 説明文とタグが最も悪い結果となった。タイトルとタグは比較的クラスタリングしやすい特徴を持ち，逆に説明文は特徴が小さいと考えられる。また，(i) タグのみと比較して，(iv) タイトルと説明文とタグの結果が悪いため，説明文の比重がタイトルやタグより大きいことが推察される。

タグ毎に見ていくと表 3 と同様に，(i)〜(iv) いずれの 6-means の結果でも Minecraft，政治の 2 タグでは良好なクラスタリングが行えたのに対して，音楽，歌ってみた，踊ってみたの 3 タグでは対照的にクラスタの混同が見られた。また，FPS は一部が Minecraft のクラスタと混同が見られた。前者 2 タグでは他タグと分別可能な大きな特徴があるのに対して，後者 3 タグでは特徴が小さかったと考えられる。

6 おわりに

本論文ではニコニコ動画の特定の 6 タグを含む動画についてニコニコ大百科の単語の分散表現を用いてクラスタリングを行なった。比較的良好な結果が得られたタグもあったが，一部のタグでは他のタグとの混同が見られるなど，まだ改善の余地がある。今後，Doc2Vec を用いた動画メタデータのみを文章として見なした場合のクラスタリングや，Doc2Vec を用いたニコニコ大百科の記事を分散表現を利用したクラスタリングを検討している。

謝辞

本研究は JSPS 科研費 15K00451 の助成を受けたものです。

参考文献

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

[2] R. Řehůřek and P. Sojka : Software Framework for Topic Modelling with Large Corpora, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Framework pp. 45–50, 2010.

[3] L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

[4] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[5] K. Kamihata and E. Ito : A quantitative contents diversity analysis on a consumer generated media site, Proc. of AROB 21st 2016 (The Twenty-First International Symposium on Artificial Life and Robotics 2016), pp. 436–440, 2016.

(i) タグのみ

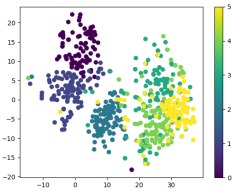

図 2: 元データ

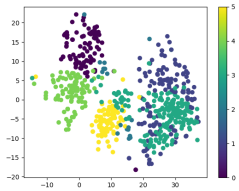

図 3: t-SNE 6-means

(ii) タイトルとタグ

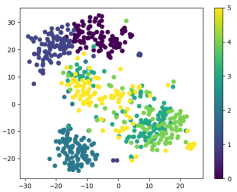

図 4: 元データ

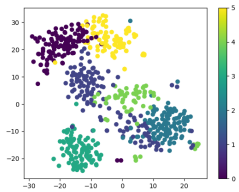

図 5: t-SNE 6-means

(iii) 説明文とタグ

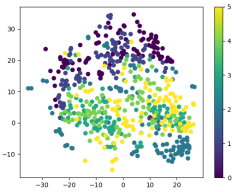

図 6: 元データ

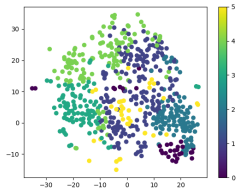

図 7: t-SNE 6-means

(iv) タイトルと説明文とタグ全て

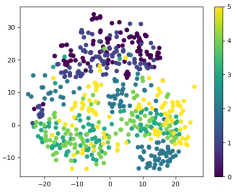

図 8: 元データ

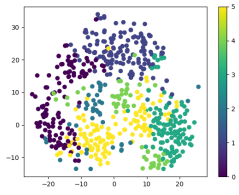

図 9: t-SNE 6-means