1T-09

SNS連動型デジタルサイネージにおける文書間距離を用いた記事掲載自動判定機構

長島悠貴† 伊藤裕二†1 小坂弘史†2 湯瀬裕昭† 渡邉貴之†
静岡県立大学† (株)メディア・ミックス静岡†1 スカパーJSAT(株)†2

1 はじめに

我々は，従来から大学向けに開発したサイネージシステムを公共空間などに設置するにあたり，サイネージの設置費や維持費を賄うための「CM再生機能」の追加など一般化を行なってきた[1]. また，災害時でもデジタルサイネージを活用し災害情報などを発信するための対応として，衛星回線を用いた通信を可能とする機能を追加し，「防サイネージ」として商業施設などに設置を進めている.

防サイネージの機能の一つに Twitter 投稿表示機能がある．この機能は，管理者がサイネージ上に掲載したいツイートを管理者の Twitter アカウントを用いてリツイートすることでサイネージ上にツイートが掲載される機能である．しかし，既存の Twitter 投稿表示機能は，サイネージ上に掲載するツイートを管理者が手動で選択しており設置箇所の増大によって管理者の管理コストも増大するという課題がある.

そこで我々は文献[2][3]において機械学習を用いて過去にサイネージ上に掲載したツイートを学習し，自動で掲載するツイートを選択する掲載記事自動判定機構を提案・評価している．分類器としてナイーブベイズ・Random Forest・SVM を使用しており，評価実験の結果精度は68%程度であった.

本研究では，先行研究において評価を行った掲載記事自動判定機構の改良を行い評価する．また，掲載ツイートを判定する際にツイートに添付された画像を Google Cloud Vision API[4]を用いて解析し，得られた結果をツイートに付与した際に精度が向上するかどうかの評価も行う.

2 関連研究

文献[5]において Kunser らは新たな文書間距離の計算手法として WMD(Word Mover's Distance)を提案している．WMD は，BoW(Bag-of-Words)が抱えていた共通語が少ない場合に意味的な類似度を測ることが困難であるという課題を別のアプローチで解決しようというものである．WMD では Word2Vec[6]を利用し，文書間距離を求める.

Word2Vec とは，単語の意味や文法を捉えるために単語をベクトル表現化して次元圧縮を行う手法である．Word2Vec は，単語ベクトルを演算処理まで行えるよう単語の分散表現として 200〜1000 次元程度のベクトルにまとめ，要素ごとではなく単語の定義によってベクトル化を行う.

WMD は文書 A,B が存在した時， A,B の語同士を対応づけることで，A を B に変換するとき対応づけの変換コストが最も低い場合の変換コストの和を文書間距離と考える．単語 i を単語 j に対応づける変換コストを式(1)のように表す.

$$c(i,\ j) = \| x_i - x_j \| \qquad (1)$$

式(1)のx_i, x_jは Word2Vec で得られた意味類似度を捉えた分散表現ベクトルである．これによって得られる変換コストの総和を求めれば文書間距離を求められることになる．しかし，式(1)のような計算は意味語の数が同じで，各単語の意味的にもほぼ 1 対 1 に結び付けられるものに対して行えることである．そこでこれを最適化問題として考えると次の式(2)のようになる.

$$\min_{T \geq 0} \sum_{i,j=1}^{n} T_{i,j} c(i,j) \qquad (2)$$

T は行列であり，各行のベクトルは，元の文書中の各単語に対応している．式(2)を解くことで WMD の計算を行う.

また，Kunser らは WMD の評価実験を行っている．BBC スポーツの記事など膨大なデータセットを用いて BoW，TF-IDF，CCG など 8 種類を K 近傍法による各分類器で生じたエラー数で比較を行った．結果として WMD は従来手法より良好な結果が出ている．特に Twitter のような単語数の比較的低い文書からなるデータセットでは，従来手法より特に良好な結果が得られている.

そこで本研究では，この評価実験の結果を踏まえ，WMD を利用してツイートの掲載判定を行う.

A method for automatically determining whether to display articles based on document distances for SNS linked digital signage

† Yuki Nagashima, Hiroaki Yuze, Takayuki Watanabe, School of Management and Information, University of Shizuoka.
†1 Yuji Ito, Media Mix Shizuoka Co., Ltd.
†2 Hiroshi Kosaka, SKY Perfect JSAT Corporation.

3　掲載記事自動判定機構

本研究では，Doc2Vec で各文書の分散表現ベクトルモデルを構築する．

Doc2Vec は任意の長さの文書をベクトル化する技術であり，Word2Vec を単語レベルではなく文や文書といった．任意の長さを扱えるように拡張したものである．Doc2Vec には PV-DM モデルと DBoW の 2 つのアルゴリズムがある．PV-DM は Word2Vec の CBoW[6]を拡張したものである．PV-DM では，CBoW において単語列のみだった入力層が図 1 のように単語列に文書 ID を付与したものになっている．DBoW モデルは，Word2Vec の Skip-gram[6]モデルを拡張したものである．DBoW では，図 2 のように入力層が文書 ID となっていることである．

Doc2Vec のパラメータは予備実験で最適なパラメータの探索を行った．探索結果を表 1 に示す．

Doc2Vec で得られたベクトルモデルから WMD を用いて学習データと新たに与えた未知データの文書間距離の計算を行う．未知データのツイートに対する文書間距離が一番近い学習データのツイートがサイネージ上に掲載されたかどうか参照を行う．掲載されていた場合には未知データの投稿を掲載，そうでなかった場合には不掲載と判断する．

ツイートに添付された画像を解析した結果を付与したデータについても同様に判定を行う．

4　性能評価

本研究では，道の駅「遠野風の丘」に設置されているサイネージ管理者の Twitter アカウントを用いて，ある観光施設のユーザアカウント A について評価を行った．同様に，ツイートに添付された画像を解析した結果を文書に付与したものについても WMD による判定を行う．自動判定機構で得られた結果と，未知データの投稿が実際にサイネージ上に掲載されたか否かを比較し評価を行う．同時に，画像の分析結果を文書に付与した場合の判定結果についても評価指標に基づいて評価を行う．評価指標として，正答率，適合率，再現率，および F 値を求める．

データセットとして，アカウント A が 2017 年 4 月 27 日から 2017 年 12 月 17 日までに投稿したツイート 747 件を利用している．そのうち画像が添付されているツイートは 563 件であり，画像総数は 863 枚であった．Doc2Vec の学習データとして 619 件を利用した．128 件を未知データとして利用し計算を行った．実験結果を表 1 に示す．

表 1 からデータに合わせたパラメータの調整で精度の向上が確認できた．また，先行研究で行った自動判定[3]と比較して精度が向上した．

5　まとめ

本研究では，先行研究において評価を行った掲載記事自動判定機構について，Doc2Vec と WMD の組み合わせによる精度の評価を行った．また，判定を行う際にツイートに添付された画像について解析を行い，その結果をツイートに付与した場合の精度についても評価を行った．今後は，WMD を機械学習に拡張した研究[7]も発表されているため本研究への活用を検討したい．

参考文献

[1]. 長島悠貴, 工藤直哉, 伊藤裕二, 小坂弘史, 湯瀬裕昭, 渡邉貴之, “観光防災対応デジタルサイネージシステムの開発”, 第 3 回とうかい観光情報学研究発表会, 2017 年 3 月

[2]. 長島悠貴, 工藤直哉, 伊藤裕二, 小坂弘史, 湯瀬裕昭, 渡邉貴之, ”CGM 連携機能を持ったデジタルサイネージにおける投稿自動判別システム”, 情報処理学会第 79 回全国大会, 2017 年 3 月

[3]. 長島悠貴, 伊藤裕二, 小坂弘史, 湯瀬裕昭, 渡邉貴之, “デジタルサイネージにおける投稿判別手法の比較”, 観光情報学会第 14 回全国大会, 2017 年 7 月

[4]. Google Cloud Vision API, https://cloud.google.com/vision/?hl=ja

[5]. M.J.Kusner, Y.Sun, N.I.Kolkin, K.Q.Weinberger, “From Word Embeddings To Document Distances”, International Conference on Machine Learning, 2015

[6]. T.Mikolov, K.Chen, G.Corrado, J.Dean, “Efficient Estimation of Word Representations in Vector Space”, arXiv preprint arXiv:1301.3781, pp. 1–12, 2013.

[7]. H.Gao, M.J.Kusner, K.Q.Weinberger, S.Fei, “Supervised Word Mover’s Distance.”, Advances in Neural Information Processing Systems. 2016.

表 1 使用パラメータ

タイプ	alpha	size	window	iter	min_count
1	0.025	300	15	400	1
2	0.025	300	10	10	10

表 2 実験結果

タイプ	データセット	正答率	適合率	再現率	F 値
1	テキストのみ	71.1%	0.6057	0.6540	0.6105
	画像解析結果付き	70.3%	0.5126	0.5134	0.5128
2	テキストのみ	70.3%	0.4965	0.4965	0.4965
	画像解析結果付き	73.4%	0.5349	0.5325	0.5334

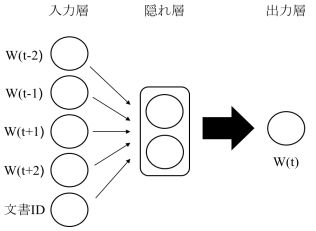

図 1:PV-DM モデル

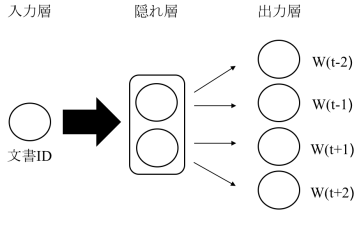

図 2:DBoW モデル