ARG WI2 No.11, 2017
WI2-2017-30

画像-テキスト間対応を利用したスポット情報拡張の定量評価とその多言語対応の検討

有山 俊一郎[a]　　延原 肇[b]

筑波大学大学院 システム情報工学研究科 知能機能システム専攻

a) *ariyama@cmu.iit.tsukuba.ac.jp*　b) *nobuhara@iit.tsukuba.ac.jp*

概要　スポット情報推薦サービスにおいて，スポットに紐づくテキスト情報がないため推薦候補として選択されない問題を解決するために，画像-テキスト間写像を利用したスポット情報拡張手法が提案されている．この手法には，いくつか調整可能なパラメータがあるが，この手法によって拡張した情報を定量的に評価できないため，パラメータの最適化が困難であった．そこで本研究では，スポット間においてそれぞれに紐付けられた画像と単語の相似性を比較することで，スポットに付与した単語がふさわしいものであるかを，画像間距離空間と単語間距離空間を融合し，定量的に評価する指標を提案し，スポット情報拡張の最適化を実現する．また，東京，ニューヨーク周辺のスポット，それぞれ 1000 スポットを利用した評価実験を行い，提案する定量指標およびそれに基づくスポット情報の一拡張を示す．

キーワード　画像-テキスト間対応，Bag of Visual Words，Word2vec，スポット推薦，多言語対応

1　はじめに

現在，GPS の利用は一般的なものとなっており，スマートフォンを通した位置情報の把握は容易になっている．著者らの研究室においても，企業との共同研究としてライフログアプリケーション FourDiary[1] を研究開発し，リリースしている．このアプリケーションは，ユーザが訪れた場所を GPS 情報に基づき自動的に記録していき，日記のような形で振り返ることができるアプリケーションとなっている．また，図 1 のようにアプリケーション上で，訪れた場所に関する，他のソーシャル・ネットワーキング・サービス (SNS) から得られた情報を閲覧することができるほか，ユーザの行動履歴からこれまでに訪れていない新たなスポットを推薦する機能の開発を行っている．

図 1　FourDiary の画面例
(左:行動履歴表示例 右:スポット関連情報表示例)

本研究に関連する，位置情報ビジネスの市場規模は 2012 年時点では 19.8 兆円であるが，2020 年には 62.2 兆円と，今後注目される市場になると予想されている [2]．SNS が普及し，多くの情報を手軽に受け取ることができる現在，情報を取捨選択した上で高精度にユーザに提示するという技術は，ユーザの負担を軽減するために重要である．こうした中で，ソーシャル・ネットワークを利用した推薦システムの研究が多くの研究者によって行われている [3] [4].

本研究では，このような SNS および位置情報サービスの中でも，FourDiary の機能にもあるようなスポット情報推薦サービスに注目し，ユーザに対して精度の高い情報推薦を目指している．そのためにはどのようなデータを利用して情報推薦を行うか見極める必要がある．本研究では，その事前調査としてユーザ投稿型スポット情報推薦サービスである Foursquare[5] に注目し，このようなサービスを利用するユーザがどのような情報 (画像やテキストなど) を多く共有しているかを国内外の約 20 万強のスポットについて調査した．その結果が表 1 である．

表 1 各地域におけるスポットデータ

	調査スポット数	平均画像数	平均投稿数
東京 23 区	130359	23.6	1.1
ニューヨーク	76230	1.7	0.7
パリ	60491	4.3	2.5

表 1 において平均画像数，平均投稿数はそれぞれユーザがスポットに投稿している画像，レビューの平均数である．この調査の結果，まず東京 23 区においては，投稿情報として画像情報がテキスト情報よりも多く投稿されていることがわかる．また，この状況は日本だけではなく，国外のニューヨークやパリなどにおいても同様で

あることが判明した．これらの画像とテキストの投稿割合をグラフに表したのが図 2 である．

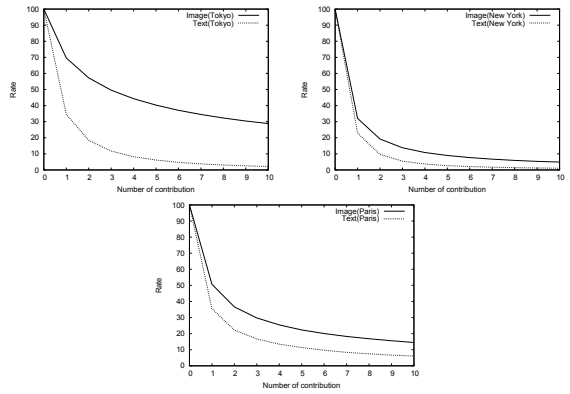

図 2　スポットに対する画像とテキストの投稿割合
(左上:東京 右上:ニューヨーク 下:パリ
実線は画像，点線はテキストを表す)

グラフの横軸の数字は投稿件数 n であり，縦軸は n 件以上の投稿が存在するスポットの割合を表している．このグラフからも画像に対してテキストの投稿数が少なくなっていることが確認できる．以上の事前調査より，人はスポットに対して，手間のかかるテキストレビューを投稿するよりも，手軽な画像を投稿する傾向にあると言える．これらのことから，ユーザ投稿を利用したスポット情報推薦サービスにおいてはテキストよりも画像を用いた情報推薦を行うことが好ましいと考えられる．画像を中心としたユーザプロファイリングを行う研究として画像-テキスト間写像についての研究 [6] [7] が提案されており，画像の類似度を用いて情報推薦を行う一方で，ユーザに対しては特徴語として単語を示すことで，推薦を受けるユーザにとって本当に嗜好にあった内容であるか判断しづらいという問題点を解決している．この手法では，Scale Invariant Feature Transform(SIFT) 特徴量 [8] による Bag of Visual Words を用いてスポットに投稿された画像を表現し，画像間の類似度を算出できるようにしている．これにより，テキスト (特徴語) を持たない画像と一番類似度の高い，特徴語を持つ画像を選択し，その画像が持っている特徴語を付与することで，特徴語を持たない画像に対しても特徴語を割り当てている．また，従来手法の構成ではスポットに付与された特徴語に関して，スポットにふさわしい特徴語が与えられたかどうかを定量的に評価できず，手法の評価には主観評価実験を行うしかないが，主観評価実験を行うには実施コストが高く，高頻度の実施は困難である．

そこで本研究では，スポット間においてそれぞれに紐付けられた画像と単語の類似性を比較することで，スポットに付与した単語がふさわしいものであるかを定量的に評価する指標を提案し，スポット情報拡張の最適化を実現する．具体的には，1000 スポットの画像特徴-特徴語データを元に Word2vec を利用して 200 スポットの画像特徴データに特徴語を付与し，実際に提案指標を用いて評価を行う．また，英語をはじめとした多言語上でのスポット情報拡張の最適化についても検討を行う．

2　画像-テキスト間写像に関する従来研究 [7]

2.1　従来手法の概要

画像-テキスト間写像を反映した従来のシステムの流れを図 3 に示す．従来手法では，まず図 3 の BoVW と特徴語の対応を扱うデータベースを生成する．具体的には，スポットに投稿されている画像から Bag of Visual Words (BoVW)[9] を抽出し，一方でそれらのスポットに投稿されているテキストに関しても名詞のみを特徴語として抽出し，紐付けた状態でデータベースに保存する．

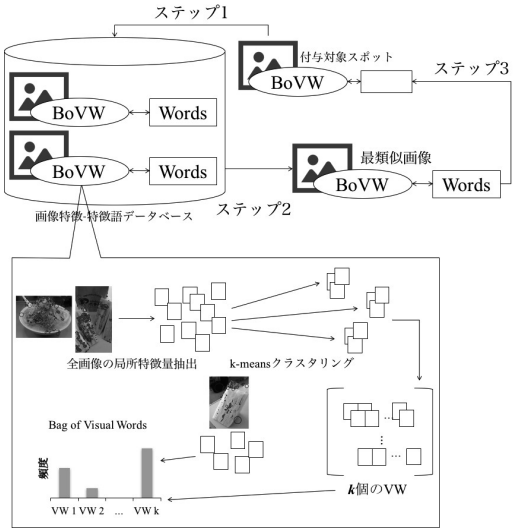

図 3 従来手法の概要 (上部) と BoVW の構成概要 (下部)

このデータベースの各 BoVW と入力画像を比較する (ステップ 1) ことで，特徴語を持たない入力画像に対して最類似画像を選択する (ステップ 2)．そして，最も類似する画像が持っている特徴語群から特徴語を，特徴語を持たないスポットへ付与する (ステップ 3)．

2.2　画像特徴-特徴語データベースの生成

従来手法では，BoVW を用いて画像の類似度を評価している．BoVW モデルでは，まず，局所記述子を用いて画像群全体から Visual Words と呼ばれる記述子の典型例を用意する．この Visual Words を用いて各画像を表現しなおすことで，画像間の比較を可能とし，類似度が計算できる．また，局所記述子としては SIFT 特徴を用いている (図 4)．

前述した通り，BoVW を用いた類似度の評価には Visual Words を用意する必要がある．まず，全ての画像

図 4　SIFT 特徴量の抽出例
(左:原画像 右:SIFT 特徴量抽出の様子)

について SIFT 特徴量を抽出する．1 つの SIFT 特徴量は 128 次元のベクトルで構成されており，1 つの画像からは複数の SIFT 特徴量を得ることができる．そして，全ての SIFT 特徴量を k-means クラスタリングにより k 個のクラスタに分け，これら k 個のクラスタのセントロイドをそれぞれ Visual Words とする．また，本研究で行う実験では，すべて $k = 1000$ としている．

BoVW は k 個の Visual Words を次元とするヒストグラムで表現される．これは，k 個の Visual Words を算出したあと，1 つの画像の全ての SIFT 特徴量について最も近い Visual Words を選択していくことでヒストグラムを構成する．これにより，1 つの画像を k 次元の特徴ベクトルで表現することが可能となる．こうして表現された特徴ベクトルは，画像が似ていればベクトルの構成も似ていると考えられ，ベクトルを比較することで画像同士の特徴を比較することができる．

2.3　特徴語付与

従来手法の目的は，特徴語を持たないスポットに対して特徴語を付与することで，情報推薦の精度を向上させることである．特徴語付与の手順は以下のように行うこととする (図 3 上部).

1. 特徴語を持たない，画像のみのスポットを選択
2. 1 のスポットの画像の BoVW とデータベース内の BoVW を比較し，最も近い画像を選択
3. 2 の画像に紐付けられている特徴語群を，新たな特徴語として 1 のスポットに付与

従来手法では，特徴語を持たないスポットに対して特徴語を付与することができたが，特徴語付与の手法の有効性についての評価は，主観評価に依存するため実施コストが高く，手軽にかつ高頻度に行うことができない．よって，キーワード付与手法の逐次改良や，ハイパラメータの調整を行った効果などを検証することが難しい．そこで本研究では，BoVW を利用した画像間距離と Word2vec を利用した単語間距離の 2 つから特徴語付与手法を評価する関数を設定し，特徴語付与を行った手法に対して定量的な評価を可能とする．

3　キーワード付与手法の定量評価およびそれに基づく改良

3.1　キーワード付与手法の定量評価指標の提案

本研究で提案する手法では，画像間距離空間と単語間距離空間を構成し，それらに基づく定量評価を行う．画像間距離は，2.2 で説明した BoVW によって構成できるが，単語間距離は何らかの方法によって，各単語を数値化し表現しなければならない．本研究では，この方法として Word2vec[10] を利用する．Word2vec は，テキスト処理を行うニューラルネットワークであり，コーパスを元に単語の特徴量ベクトルを生成することができる．本研究では 200 次元のベクトルを構成し，この特徴量ベクトルを利用することで，Word2vec は与えられた単語の意味の推測を行うことができ，さらに，ある単語と他の単語との関連性を求めることができる．これらによって，短期間での改良による検証が行いやすく，また，検証自体にかかる時間も短縮できる．この評価指標の概要を図 5 に示す.

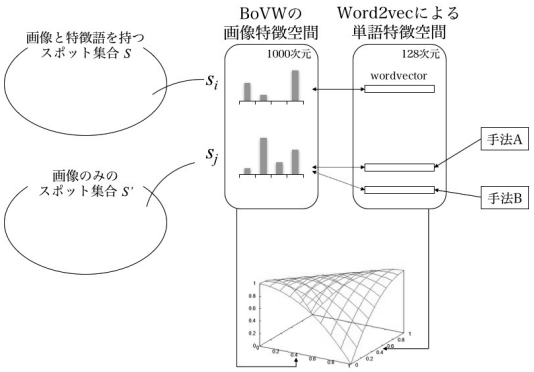

図 5 提案する評価指標の概要

提案する評価指標の定義を以下に述べる．まず，スポット集合 S に属するスポット s_i と s_j において，そのスポット間の画像間距離を $I(s_i, s_j)$，特徴語による単語間距離を $W(s_i, s_j)$ としたとき，それぞれ，

$$I(s_i, s_j) = \sqrt{\sum_{k=1}^{n} \left(H_{s_i}(k) - H_{s_j}(k) \right)^2} \qquad (1)$$

$$W(s_i, s_j) = 1 - \frac{\sum_{k=1}^{n_{s_i}} \sum_{l=1}^{n_{s_j}} w\left(s_i(k), s_j(l)\right)}{n_{s_i} n_{s_j}} \qquad (2)$$

で求められる値を最小値が 0，最大値が 1 となるよう正規化したものとする．ここで $H_{s_i}(k)$ はスポット s_i に

おける画像特徴のヒストグラムの k 番目の次元を表し，BoVW により生成したヒストグラム間のユークリッド距離を求めることとなる．また，$w(s_i(k), s_j(l))$ は，スポット s_i における k 番目の特徴語とスポット s_j における l 番目の特徴語の類似度を Word2vec により算出する関数である．つまり，2 スポット間の全ての特徴語同士の類似度平均を求めることとなる．また，単語類似度は 1 に近いほど類似していることを示すが，評価関数に適用するために値が小さいほど評価が高くなるよう，1 から引いている．

以上の画像間距離および単語間距離を用いて，最終的なスポット間の類似度の評価値を，

$$Score_{ij} = 1 - (I_{ij} - W_{ij})^2 \tag{3}$$

で求めることとする．評価関数については，他の形状も考えられるが，最も基本的な形としてこのような定義を行う．この評価関数の概形を図 6 に示す．

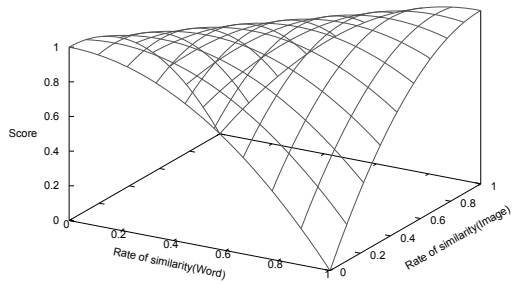

図 6 評価関数の概形

画像間距離及び単語間距離は 0 に近いほどお互いが似ていることを表し，画像間距離と単語間距離が共に近い，もしくは共に遠い時に高く評価される形になっている．この評価値をスポット集合 S に属する全てのスポットの組み合わせで算出し，その合計を特徴語付与手法の評価値とする．

3.2 キーワード付与手法の改良

特徴語付与を行うにあたって，まず，次のような手順で基準語を選定することとする．

1. 特徴語を持たない，画像のみのスポットを選択
2. 1 のスポットの画像の BoVW とデータベース内の BoVW を比較し，最も近い画像を選択
3. 2 の画像に紐付けられている特徴語群の全ての単語に対して，特徴語群に属する他の単語との類似度を計算し，平均値を算出
4. 最も平均値の高い単語を基準語として選定

Word2vec を用いて関連性の高い単語を抽出するためには，どの単語を基準に関連性を求めるか決める必要があるため，基準語の選定を行うこととした．また，単語類似度の平均値を利用しているのは，類似度平均値の高い特徴語はその特徴語群の中で他のどの単語とも関連性があると考えられるため，特徴語群を代表する特徴語と言えるからである．

3.2.1 Word2vec による特徴語付与

基準語の選定によって 1 つの単語を選び出した後は，Word2vec を利用して基準語と関連性の高い単語を特徴語の候補として抽出する．そして，関連性の高い単語から順に規定数に達するまで特徴語として付与を行うこととする．

4 東京都心のスポットを対象とした評価実験

提案手法の有用性を確認するために，定義した評価関数を用いた評価実験を行った．評価実験を行うに当たって，3.1 で定義した評価指標について，具体的なスポットについてどのように評価されるか例を挙げる．図 7 と表 2 を基準として，図 8 および表 3 は評価値が高い例で，画像と特徴語が共に似ていると判断されたものである．逆に図 9 および表 4 は評価値が低い例で，画像は似ているが特徴語は似ていないと判断されたものである．また，表 5 はそれぞれの実際の評価値である．

図 7 基準となる画像例

表 2 基準となる特徴語例

特徴語
定食

図 8 評価値が高い画像例

表 3 評価値が高い特徴語例

特徴語
御飯
定食

図 9 評価値が低い画像例

表 4 評価値が低い特徴語例

特徴語
イス
数
テイクアウト
前提

表 5 実際の評価値

	画像類似度	単語類似度	評価値
高評価例	0.0078	3.331×10^{-16}	0.9999
低評価例	0.0087	0.6623	0.5728

方法としては，従来研究による特徴語付与を行ったデータベースと提案手法による特徴語付与を行ったデータベースを用意し，データベース内の全てのスポットデータに対して評価関数を用いて評価し，その評価値の合計を比較するというものである．用いたスポットは東京 23 区とその周辺に存在する 1200 スポット，そのうち 200 スポットが特徴語を持たないスポットであり，1000 スポットの画像特徴と特徴語から残り 200 スポットの特徴語を付与する形で実験を行った．新たに付与する特徴語は 1 つのスポットにつき最大で 5 つまでとし，それぞれの手法についての評価値は，各スポットに対する評価値の合計としている．実験の結果得られた評価値を表 6 に示す.

表 6 実験により得られた評価値 (東京都心)

	評価値
従来手法	368697.343
提案手法	375849.609

結果としては，従来研究による特徴語付与よりも約 2%評価値が向上した．この 2%の向上について，具体的に特徴語の変化がどのように現れたかを示したのが表 7 と表 8 であり，それぞれの平均値が表 9 である．また，このスポットを表す画像が図 10 である.

図 10 特徴語が付与されていないスポット画像例

表 7 具体的なスポット例 (従来手法)

	日替わり	定食	魚	野菜	作り
日替わり	1	0.294	-0.019	0.126	0.105
定食	0.294	1	0.360	0.551	0.199
魚	-0.019	0.360	1	0.654	0.330
野菜	0.126	0.551	0.654	1	0.426
作り	0.105	0.199	0.330	0.426	1

表 8 具体的なスポット例 (提案手法)

	野菜	果物	食材	豚肉	魚介類
野菜	1	0.877	0.826	0.819	0.814
果物	0.877	1	0.811	0.832	0.849
食材	0.826	0.811	1	0.808	0.778
豚肉	0.819	0.832	0.808	1	0.833
魚介類	0.814	0.849	0.778	0.833	1

表 9　図 10 に示すスポット画像に対する特徴語付与の評価値比較

	平均値
従来手法	0.442
提案手法	0.860

結果の表を見ると，提案手法の方が，付与された特徴語間の類似度が高い値となっており，関連性の高い単語を特徴語として付与できていることがわかる．しかし，特徴語間の類似度は高くなったものの，評価値は大きく向上しなかったことも確認できる．その理由として，提案手法では最も似ている画像を元に特徴語の付与を行っているが，どの程度似ているかという画像間の距離は利用していないということが挙げられる．つまり，評価関数の値は画像間距離と単語間距離の値が近いほど高い評価となるため，画像間距離を元にその値と近い単語間距離となる単語を中心に特徴語として付与することで改善が行えると考えられる.

5　多言語対応に関する検討

今回の提案手法が主に日本語で記述されている東京都内のスポットだけでなく，日本国外のスポットでも有用であることを示すために，ニューヨーク市周辺のスポットを利用して同様の実験を行った．スポット数は第 4 章と同じく 1200 スポットで，そのうち 200 スポットが特徴語を持たない状態で開始した．まず，実験の結果得られた評価値を表 10 に示す.

表 10 実験により得られた評価値 (ニューヨーク市周辺)

	評価値
従来手法	310013.383
提案手法	311478.289

結果としては，従来手法，提案手法共にほとんど変わらない値となった．提案手法の評価値が従来手法に比べて大きく変化しなかった原因としては，実験対象として選んだ1200スポットの選び方にあると考えられる．本研究の実験では，対象となるスポットをその地域のスポットから無作為に選んでいるが，1200スポットという数はその地域に存在するスポットの数に比べてかなり小さな数である．そのため，その地域に存在する様々なジャンルのスポット全てを網羅しているとは考えにくい．本研究では類似している画像を元に特徴語の付与を行うため，同じジャンルのスポットがデータベース内に存在しないと，最も似ている画像の類似度が著しく落ちてしまい，それに基づいて付与される特徴語も付与先の画像から離れたものになってしまう．したがって今後は，実験対象のスポットを選ぶ際にある程度ジャンルを絞ってスポットの選択を行い，再度同様の実験を行うことで評価値が向上するか確認し，有用性の評価を行いたい．

6 まとめ

本研究では，SNSにおいてテキスト情報よりも画像情報の方が共有されやすいという特徴を元に，テキスト情報が少ない場合でも情報推薦を行えるように，画像-テキスト間対応を利用したスポット情報の拡張手法を提案した．また，スポット情報の拡張手法を評価する方法が主観評価実験しかなく，実施コストが高いという課題を解決するために，新たに定量評価指標を提案した．これにより，手軽かつ高頻度な評価が行えるようになり，特徴語付与手法の逐次改良が可能になると考えられる．

提案指標を利用したスポット情報拡張手法の評価実験では，従来手法と提案手法を実際に提案指標で評価を行い，比較した．結果としては従来手法と提案手法の差はわずかな値であったが，提案手法による精度の高い特徴語の付与が行えていることを確認できた．また，英語圏に存在するスポットに対しても評価実験を行い，日本語圏と同様の精度が得られるか検証した．その結果，こちらも提案手法が従来手法を上回る結果となったが，その差はわずかであった．今後は，実験を行う際のスポットをジャンルに絞って選ぶことで，類似画像にヒットしやすくすることで評価値に変化が見られないか検証するほか，多言語での特徴語付与についてさらなる分析を行っていく予定である．

参考文献

[1] FourDiary, available from (fourdiary.com).
[2] 総務省：G 空間× ICT 推進会議報告書，2013.
[3] Allison J.B. Chaney, David M. Blei and Tina Eliassi-Rad: A Probabilistic Model for Using Social Networks in Personalized Item Recommendation, RecSys '15 Proceedings of the 9th ACM Conference on Recommender Systems.
[4] 澤井里枝，有安香子，藤沢寛，金次保明：SNSを利用した協調フィルタリングによる番組推薦手法，研究報告データベースシステム (DBS)，Vol. 2010-DBS-151, No. 43, pp. 1-8, 2010.
[5] Foursquare, available from (foursquare.com).
[6] 有山俊一郎，延原肇：単語間類似度を考慮した画像-テキスト間写像構成手法の検討とその位置情報サービスへの応用，第9回 WI2 研究会，2016.
[7] 大東祐太，有山俊一郎，延原肇：位置情報 SNS 上の画像-テキスト間対応を利用したユーザ嗜好抽出と推薦スポット候補拡張，情報処理学会論文誌，Vol. 58, No. 12, pp. 1-9, Dec. 2017.
[8] David G. Lowe: Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, Vol. 60, No.2, pp. 91-110, 2004.
[9] Csurka, G., Bray, C., Dance, C. and Fan, L.: Visual categorization with bags of keypoints, in Proc. of ECCV Workshop on Statistical Learning in Computer Vision, pp. 59-74, 2004.
[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean: Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing Systems 26 (NIPS 2013).