WI2 *Web Intelligence and Interaction*

ARG Web インテリジェンスとインタラクション研究会

ARG SIG-WI2

第 16 回研究会 予稿集

WI2-2020-1～24

WI2-2020-T1

2020 年 11 月 27 日～28 日

オンライン

ARG Web インテリジェンスとインタラクション研究会予稿集目次

（第 16 回研究会）

CONTENTS

招待パネル

セッション3：テキスト処理

11 月 28 日（土）

セッション 4：機械学習と推薦システム

セッション 5：SNS

セッション 6：学術情報

Note: All rights are reserved and copyright of this manuscript belongs to the authors. This manuscript has been published without reviewing and editing as received from the authors: posting the manuscript to ARG WI2 does not prevent future submissions to any journals or conferences with proceedings.

ARG WI2 No.16, 2020
WI2-2020-01

運転者の視線を考慮した運転難易度評価の検討と特定地域の分析

福島 秀敏[†,a]　荒木 徹也[††,b]　石川 博[†,c]

† 東京都立大学大学院システムデザイン研究科情報科学域　†† 群馬大学理工学部電子情報理工学科

a) *fukushima-hidetoshi@ed.tmu.ac.jp*　b) *tetsuya.araki@gunma-u.ac.jp*　c) *ishikawa-hiroshi@tmu.ac.jp*

概要　自動車をめぐる問題の一つに運転の難しさがある．道路を運転する難しさを定量的に表すことができれば，交通事故の分析や，運転経路の探索など様々な研究に役立てることができる．運転は，認知，判断，操作の連続によって行われており，事故の要因としては，認知の失敗が最も多く挙げられている．そこで本研究では人間の認知に着目した運転難易度の評価手法を提案する．また，それを用いた特定地域の運転の難しさの可視化を行う．具体的には，Google Street View と OpenStreetMap を組み合わせて運転画像を収集し，顕著性マップを適用する．運転者の視界は中央に偏るため，適用範囲をトリミングにより変更して実験を行なった．結果，トリミングを用いた場合，計算結果とアンケート結果の間に-0.6160 の相関値が得られた．可視化結果からは，見晴らしの良い道路は運転が容易であるという考察が得られた．

キーワード　地理情報, 運転難易度, 顕著性マップ

1　はじめに

自動車をめぐる問題の１つに運転の難しさがある．運転の難しさには，運転手の熟練度や自動車の種類，道路の状況など，様々な要因が影響していると考えられる．そのため，運転の難しさを定量的に評価することは難しい．また，運転の難しさにはさまざまな捉え方がある．一般に運転は，認知，判断，操作の３要素の繰り返しであると考えられている．交通事故総合分析センターの調査 [1] では，認知にまつわる事故が一番多く挙げられている．これは，認知に失敗すれば，判断や操作にも繋がらないため，交通事故発生の要因になりやすいためであると考えられる．

近年，条件付き運転自動化が解禁された．これにより，一定の条件を満たした区間での自動運転が可能になる．そして，これらシステムの登場により，運転の難しさが軽減すると考えられる．しかし，自動運転システムはまだ完全ではなく，一定の条件を満たした区間を走る場合などに使用条件が限られる．そのため，自動運転を使用する場合，走行経路の変更などが必要になると予想される．走行経験のない道路では，システムによる運転時間や距離の予想は可能であるが，目的地までの運転の難しさがどの程度軽減されるのかを判断することは難しい．これらは自動運転に伴う経路問題であると捉えることもできる．

本研究では，運転の認知的な難しさに着目し，道路の運転難易度評価を行う．本研究を応用することで，事故の分析や経路探索の研究に繋げられると考えられる．データセットとして Google Street View[1] を利用する．Google Street View には，記録用の自動車が走行した際に収集された様々な道路上の画像が存在する．Google Street View を用いることで，任意の道路の擬似的な運転画像を収集できる．一般に自動車に関連した研究では，対象となる道路を実際に自動車で走る必要が考えられるが，これらの擬似的な運転画像から道路の運転難易度評価を行うことができれば，実験コストや被験者の負担が少なくなるため，有益であると考える．

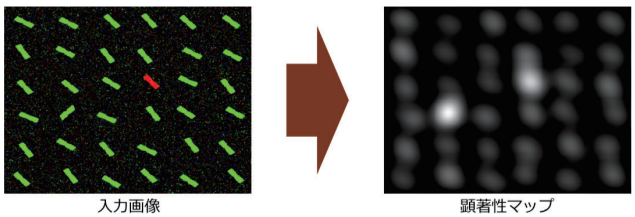


図 1 spectral-eigenPQFT

運転難易度の算出にあたり，顕著性マップを用いる．顕著性マップとは，人間の視覚の知覚モデルである．図 1 に，顕著性マップを使用した場合の例を示す．図 1 の入力画像には，斜め右下を向いた緑色の点が 34 個，それと同じ方向を向いた赤色の点と逆を向いた緑色の点がそれぞれ 1 個存在している．顕著性マップを用いることで，赤色の点と逆を向いた緑点が白く強調されているのが分かる．

本研究の先行研究 [2] では，Google Street View から人手で道路画像を集め，その道路の運転状況の難しさをアンケートした．また，顕著性マップを用いた計算指標を示し，アンケート結果と計算指標の間に中程度の相関値を得ている．本研究では，運転中の人間の視界が中心に偏ることを考慮し，計算指標の相関値向上を目指す．また，特定の範囲における Google Street View 画像の収集方法提案を行い，最も相関値の高い運転難易度の指標を用いて可視化を行う．

Copyright is held by the author(s).
The article has been published without reviewing.

[1] https://www.google.com/streetview/

2　関連研究

2.1　運転の難易度に関する研究

自動車の運転の難しさに着目した様々な研究が行われている．Zhang ら [3] は，自動運転データセットの道路の複雑さを意味記述子で要約し，SVM で 3 段階の難易度に分類した．Takeda ら [4] は，カーシミュレータと聴覚プローブを用いて，運転の難易度と楽しさの関係を示した．また，連続的なカーブよりも単体の鋭角カーブの方が，難易度と楽しさが増加することを示した．Ryan ら [5] は，人間の走行データから，自動運転にも適応できるリスクの代理測定手法を示した．Wanga ら [6] は，ブレーキ時の特性から，運転リスクを定量化する手法を提案した．これらの研究は，シミュレータや実際の走行を伴うデータの収集を行っている．本研究は，Google Street View を用いて，世界中のほぼすべての任意の場所での道路の運転難易度評価を実現することを目標にしている．

2.2　視線に関する研究

三浦 [8] は，運転中の視覚移動は，状況，車種，速度，個人によって異なり，かつ視覚的な多重課題性をおびるものであることを示唆した．Nakayasu ら [9] は，運転中の人間の眼球運動とさまざまな道路状況下での関係を評価した．また，周辺視野の視覚刺激が運転中の眼球運動に影響を与えることを示した．Miltenburg ら [10] は，熟練した運転者の方が，熟練していない運転者に比べて，簡潔に視線の停留が行われていることを示した．複雑な道路では人間は視線を中央から移動させる必要がある．したがって，運転の難しさと視界の複雑さには何らかの関係があると考えられる．

顕著性マップに関する様々な研究 [7] が行われている．視覚的顕著性は，Koch ら [11] によって提唱された．顕著性とは，人間の視覚を元にされた概念であり，人間の注意を引く可能性がどの程度あるかを表す．本研究では，スペクトルを用いた顕著性マップのモデルを用いる．Hou ら [12] は，フーリエ変換に基づいたスペクトル残差を抽出し，顕著性マップを構築する高速な手法を提案した．Guo ら [13] は，位相スペクトルを用いる手法を提案した．Schauerte ら [14] は，四元数フーリエ変換に基づくスペクトル顕著性マップに固有軸と固有角度の使用を提案した．

3　手法

3.1　運転難易度の評価

3.1.1　評価用データセット

COCADOO-COgnitive CAr Driving OpiniOns[2] をデータセットとして用いる．データセットには，Google Street View から人手で収集された 23 枚の道路の画像とそれに対するアンケートの結果が含まれている．23 枚の画像に対して，アンケートの回答者は運転状況を 5 段階の難易度で評価した．また，その状況での適切な速度 [km/h]，最小の速度 [km/h]，最大の速度 [km/h] の 3 項目の速度をそれぞれ自由に回答した．

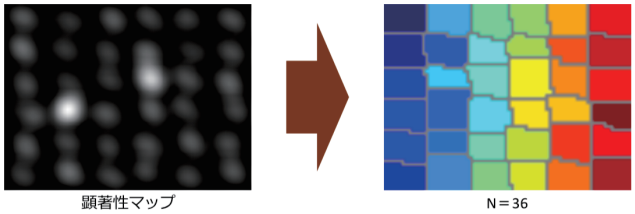


図 2 流域変換の例

3.1.2　評価指標

顕著性マップでは，視線を引き付ける場所がピークとなる．運転の認知的な難しさには，視界の複雑さが関わっていると考えられる．本研究では，以下の二つを運転難易度の指標として用いる．

- N：ピーク数

  運転者の視界から，視線を引き付ける物体の数は，運転の難しさに関係している可能性が高い．そのため，運転難易度の評価指標としてピーク数 N を求める．ピーク数 N は，出力された顕著性マップのグレースケールを反転させた後，流域変換を行うことで求める．図 2 に流域変換による指標 N の例を示す．

- A：ピーク面積

  運転手の視界から，視線を引き付ける物体の相対面積は，運転の難しさに関係している可能性が高い．そのため，運転難易度の評価指標としてピーク面積を求める．ピーク面積は，出力された顕著性マップのうち，閾値 0.9 を超えた pixel 数が画像全体を占める割合で求める．

計算指標の評価には，上記の指標とアンケート結果との相関値を用いる．相関値の計算にはピアソンの線形相関係数を用いる．本研究では，運転中の人間の視界が中心に偏ることを考慮する．そのため，データセットの画像を中央でトリミングした場合と画像全体を用いた場合でそれぞれ相関値を計算する．

3.2　運転画像の収集

擬似的な運転画像の収集にあたって，道路に沿った緯度と経度の集合が必要になる．そこで本研究では，OpenStreetMap[3] を用いる．OpenStreetMap には，Node と呼ばれる緯度と経度で構成された基本要素がある．道路

[2] http://kgwisc.aei.polsl.pl/index.php/pl/dataset/65-cocadoo

[3] https://www.openstreetmap.org/

に沿った Node を収集し，その Node の緯度と経度を擬似的な運転画像の収集箇所として用いる．

道路情報の収集にあたり，個別に選択された部分を収集することができる Overpass API[4]を用いる．クエリの実行には，ウェブベースのデータマイニングツールである overpass-turbo[5]を用いた．自動車が通行する可能性のある道路として，highway キー[6]に表 1 の値を指定した．

表 1 highway に指定した値

motorway	trunk	primary
secondary	tertiary	unclassified
residential	motorway_link	trunk_link
primary_link	secondary_link	tertiary_link
living_street	service	bus_guideway
road		

Google の提供する Street View Static API[7]を用いて，運転画像の収集を行う．収集する画像サイズとして，縦と横をそれぞれ 640pixel とした．道路には二つの進行方向が考えられるため，道路に沿った Node 集合の両端以外の Node からは両進行方向の画像を収集する．進行方向である方位角は，連続した 2 つの Node から計算を行う．

3.3 地図に可視化

特定地域における運転難易度の違いを分析するために，運転難易度評価モデルに基づいて可視化を行う．分析対象の地域は 1 辺 500m の $2,500\text{m}^2$ である．地図を 1 辺約 63m のグリッドで区切り，そのグリッド内の運転難易度評価に応じて色付けを行う．画像が収集できなかったグリッドは白色で表す．特定のグリッド内で収集された道路画像から考察を行う．

4 実験

4.1 難易度評価

手法を用いて相関値を算出した．表 2，3 にそれぞれの指標とアンケート結果との相関値を示す．それぞれの項目のうち，もっとも相関のある値を強調した．

4.2 難易度評価の考察

節 4.1 の結果を元に，運転者の視界を考慮した場合の指標 N と指標 A の相関値の考察を行う．表 2 には，本研究のベースラインとなる相関値が示されている．運転画像として，人手で収集された縦 495pixel，横 935pixel の画像を用いている．速度に関するアンケート結果との相関値は，指標 A が 3 項目で 0.58 前後の中程度の相関を示している．運転難易度では，指標 A が-0.3363 の弱い相関を示した．どのアンケート項目でも指標 A の面積を利用した手法の相関傾向が強い．

表 3 は，運転手の視線を考慮した場合の相関値である．運転画像として，中央を縦 98pixel，横 187pixel にトリミングした画像を用いている．速度に関するアンケート結果との相関値は，指標 N が 3 項目で 0.52 前後の相関を示している．これらの相関値はベースラインの値よりも低い．運転難易度の項目では，指標 N が-0.6160 の相関値を示した．これは，ベースラインの-0.3363 よりも相関傾向が強いことを示している．

提案手法では，負の相関を示したため，指標 N の値が高いほど運転が容易であり，指標 N の値が低いほど運転が難しいと考えることができる．指標 N は視線を集める物体の数であるため，画像中央に物体が集中していると運転が容易であり，画像中央に物体が少ないと容易であると言い換えることができる．トリミングにより画像の端の部分，つまり，視界の端に当たる部分に物体が存在するほど運転が難しく，中央に集中していると運転が容易であると表れた結果だと考察する．

4.3 可視化用データセット

東京タワー周辺と京都駅周辺で実験を行った．収集されたデータセットの Node 数，画像数，緯度と経度の範囲を表 4 に示す．また，収集範囲の地図を図 3，4 にそれぞれ示す．図 3，4 では，収集対象された道路を青い線で表している．

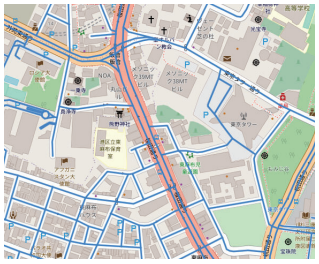


図 3 東京タワー周辺

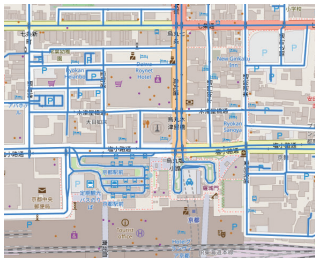


図 4 京都駅周辺

4.4 可視化

4.1 節より，最も相関傾向のあった spectral-eigenSR モデルを用いた可視化を行う．運転者の視線を考慮し，収集された画像の中央を縦 98pixel，横 187pixel にトリミングし，入力画像とした．グリッド内で収集された画像から指標 N の値を算出し，各画像の指標 N の平均値を元に色付けを行なった．図 5，6 に可視化した結果を示す．指標 N の値が高い地域は赤色で，指標 N の値が低い地域は黄色のグリッドで表されている．白色のグリッドは，画像が収集できなかった地域である．最も相関があった-0.6160 の指標 N を可視化に用いたため，赤色のグリッドで表された地域は運転が容易であり，黄色のグリッドで表された地域は運転が難しいという結果になる．

[4] http://overpass-api.de
[5] http://overpass-turbo.eu
[6] https://wiki.openstreetmap.org/wiki/JA:Key:highway
[7] https://developers.google.com/maps/documentation/streetview/

表 2 ベースラインの相関値

	適切な速度		最小の速度		最大の速度		運転難易度	
顕著性マップのモデル	N	A	N	A	N	A	N	A
spectral-DCT	-0.3804	-0.0456	-0.4137	-0.0116	-0.3522	-0.0389	0.2046	-0.0586
spectral-PFT	-0.3011	0.5299	-0.3169	0.5462	-0.2642	0.5122	0.0641	-0.3323
spectral-PFTmultiscale	-0.2342	-0.0658	-0.2089	-0.0507	-0.2785	-0.0465	-0.0655	-0.0104
spectral-PQFT	-0.4190	**0.5877**	-0.4422	**0.5907**	-0.3895	**0.5792**	0.1610	**-0.3363**
spectral-PQFTmultiscale	-0.1391	0.4016	-0.1491	0.4097	-0.1057	0.4128	-0.2248	-0.2558
spectral-QDCT	-0.1765	-0.2774	-0.1760	-0.2628	-0.1720	-0.2510	-0.1957	0.1063
spectral-QDCTmultiscale	-0.1610	-0.0995	-0.1561	-0.0841	-0.1489	-0.069	-0.2154	-0.0152
spectral-SR	-0.1240	0.2674	-0.1140	0.2661	-0.1603	0.2593	-0.0689	-0.1723
spectral-eigenPQFT	-0.1304	0.4743	-0.1454	0.4895	-0.0996	0.4660	-0.1603	-0.2497
spectral- eigenPQFTmultiscale	-0.2670	0.0393	-0.2632	0.0628	-0.2459	0.0053	-0.0942	0.0683
spectral-eigenSR	-0.1322	0.2925	-0.1250	0.3029	-0.1600	0.2733	-0.0245	-0.1606
spectral-eigenSRmultiscale	-0.2550	-0.0322	-0.2333	-0.0175	-0.2944	-0.0146	-0.0668	-0.0015

表 3 提案手法の相関値

	適切な速度		最小の速度		最大の速度		運転難易度	
顕著性マップのモデル	N	A	N	A	N	A	N	A
spectral-DCT	-0.1169	0.1723	-0.1286	0.1756	-0.1118	0.1571	-0.0304	-0.0557
spectral-PFT	0.4600	-0.0488	0.4595	-0.0450	0.4630	-0.0403	-0.4751	0.1127
spectral-PFTmultiscale	0.1022	0.2382	0.0903	0.2616	0.0959	0.2195	-0.2805	-0.3289
spectral-PQFT	0.4864	-0.0415	0.4827	-0.0318	0.4850	-0.0435	-0.5813	0.1828
spectral-PQFTmultiscale	0.1198	0.2419	0.0749	0.2613	0.1483	0.2494	-0.1915	-0.3059
spectral-QDCT	-0.1106	0.1843	-0.0943	0.2146	-0.1214	0.1664	-0.1835	-0.1614
spectral-QDCTmultiscale	-0.1324	0.2602	-0.1398	0.2589	-0.1221	0.2609	0.0194	-0.1422
spectral-SR	0.2627	0.3390	0.2558	0.3171	0.2356	0.3413	-0.5240	-0.2295
spectral-eigenPQFT	**0.5194**	0.0056	**0.5105**	0.0126	**0.5342**	0.0132	-0.5974	0.0677
spectral- eigenPQFTmultiscale	0.2594	0.2498	0.2228	0.2773	0.2807	0.2450	-0.1592	-0.3133
spectral-eigenSR	0.2179	0.3765	0.2101	0.3521	0.2104	0.3856	**-0.6160**	-0.2431
spectral-eigenSRmultiscale	0.0452	0.2749	0.0235	0.3030	0.0551	0.2591	-0.2672	-0.3326

表 4 データセット詳細

地域	Node 数	画像数	緯度の範囲	経度の範囲
東京タワー周辺	597	978	35.656250	139.740625
			35.660417	139.746875
京都駅周辺	646	1,059	34.985417	135.756250
			34.989583	135.762500

4.5 可視化結果の考察

4.4 節の結果を元に，運転難易度評価の考察を述べる．

最初に，東京タワー周辺の考察を述べる．実験対象となった東京タワー周辺は図 3 に表されている．可視化結果は，図 5 に表されている．可視化結果を表す画像上部には，グリッド内で収集された画像から計算された指標 N の平均値とグリッドの色の閾値が示されている．東京タワー周辺の地域では，特徴として，外苑通りと桜田通の交差点が挙げられる．東京タワー周辺では，5 つの地域で運転が難しく，3 つの地域で運転が難しいという結果になった．

図 5 の①番の地域で収集された画像の例を図 7 に示す．①番の画像は，永井坂の画像である．2 車線の道路であり，中央線の右側にはみ出した追い越しが禁止となっている．トリミング後の指標 N の値が 32 であり，運転が容易な道路という結果になった．N の値が高くなった要素として，画像中央の自動車が挙げられる．直線道路のため，画像中央の消失点付近に自動車が複数重なって

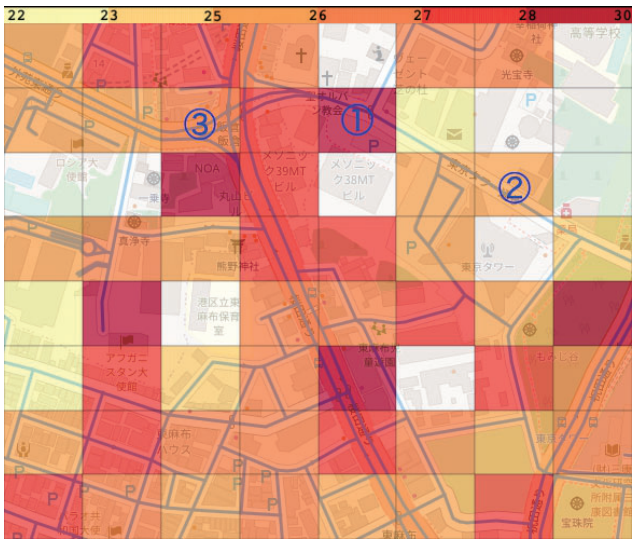


図 5 東京タワー周辺

いるのが分かる．見晴らしが良いため，遠くの建物がトリミング後に複数写っているのも N の値が高くなった要因として挙げられる．

図 5 の②番の地域で収集された画像の例を図 8 に示す．②番の画像は，東京タワー通りの画像である．1 車線の道路であり，画像右側に東京タワーが一部写っている．トリミング後の指標 N の値が 24 であり，運転するのが難しい道路ということになる．N の値が低くなった

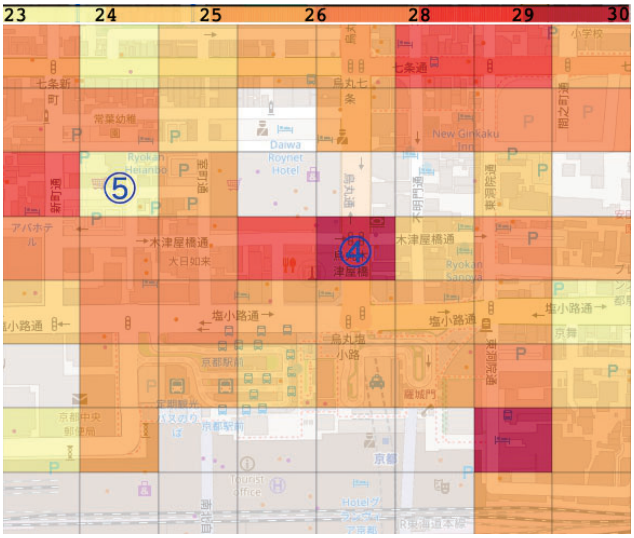
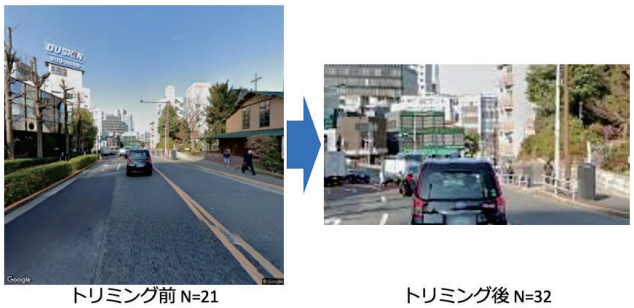


図 6 京都駅周辺

要因として，自動車の数が少なく，背景に建物などが少ないことが挙げられる．また，1 車線の見通しの悪い道路では，前を走る自動車により，さらにその前を走る自動車が隠れてしまうことが挙げられる．

図 5 の③番の地域で収集された画像の例を図 9 に示す．③番の画像は，合流によるカーブの画像である．トリミング後の指標 N の値は 29 であり，運転が容易な道路という結果になった．主観的には，合流によるカーブは運転が難しいと考えられる．N の値が高くなった要因として，見晴らしが良く，トリミング後の画像に建物が複数写ってしまったことが挙げられる．建物が少ない場合などでは，N の値が低くなる可能性があり，今後の課題としたい．

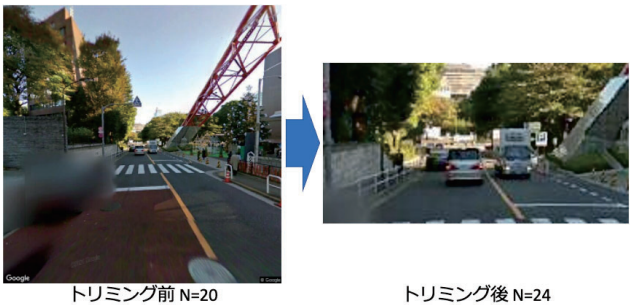


トリミング前 N=21　　トリミング後 N=32

図 7 ①番の地域で収集された画像の例

次に京都駅周辺の考察を述べる．実験対象となった京都駅周辺の範囲は図 4 に表されている．可視化結果は，6 に表されている．京都駅周辺では，特筆点として区画整理された直線道路が多いことが挙げられる．実験結果として，京都駅周辺では，2 つの地域で運転が容易であり，4 つの地域で運転が難しいという結果になった．

図 6 の④番の地域で収集された画像の例を図 10 に示す．④番の画像は，京都駅につながる見通しの良い複数車線の道路である．トリミング後の指標 N の値が 29 であり，運転が容易な道路という結果になった．N の値が高くなった要因として，反対車線を走る複数の自動車が写ったことが挙げられる．また，見晴らしが良いことで，背景の建物が写ったことなども要因として挙げられる．



トリミング前 N=20　　トリミング後 N=24

図 8 ②番の地域で収集された画像の例

トリミング前 N=25　　トリミング後 N=29

図 9 ③番の地域で収集された画像の例

図 6 の⑤番の地域で収集された画像の例を図 11 に示す．⑤番の画像は，中央通りを外れた車線のない道路である．トリミング後の指標 N の値が 16 であり，運転が難しい道路という結果になった．N の値が低くなった要素として，人や自動車が画像に写っていないことが挙げられる．道幅が狭く，トリミング後にも，両側の建物が大きく写っていることも要因として挙げられる．

収集画像の傾向から，考察を行う．指標 N が高くなる要因として，自動車や背景の建物が挙げられる．これらは複数車線の見晴らしの良い道路で発生しやすいと考えられる．また，中央でトリミングを行うことで，見晴らしのよい道路と悪い道路での違いが，はっきりと現れる可能性があることが分かった．一般に大きい道路の方が，小さい道路よりも，認知的に運転しやすいと考えられるため，この結果は妥当であると考えられる．カーブなどは，認知的な難しさとは別に操作的な難しさの要因が大きいと考えられるため，認知的な難しさだけを使用した場合，結果と齟齬が生じる可能性がある．

5　まとめと今後の課題

本研究では，運転者の視線が中央に偏りやすいことに着目した運転難易度評価を検討した．また，OpenStreetMap の道路上の Node を元に，Google Street View を用いて擬似的な運転画像を収集した．それらを用いて，

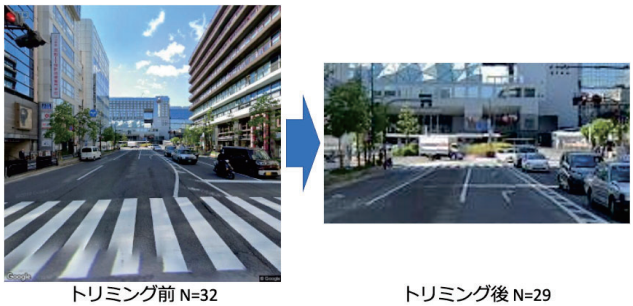


図 10 ④番の地域で収集された画像の例

図 11 ⑤番の地域で収集された画像の例

特定の地域の運転の難しさを可視化し，考察を行った．

画像を中央でトリミングした場合，データセットのアンケート結果との相関傾向が強くなり，spectral-eigenSR モデルと指標 N を用いた場合に-0.6160 の相関を示した．可視化結果の傾向として，見晴らしの良い道路では運転が容易であり，見晴らしの悪い道路では運転が難しいという結果になった．これは，認知に基づいた運転の難しさを扱ったためであると考えられる．一般に，カーブなどの操作的な難しさが高いと考えられる道路では，認知的な難しさとは別の指標が必要になると考えられるため，今後の課題としたい．

他の課題としては，最適な道路画像の大きさを調べることが挙げられる．本研究では，画像を縦 98pixel，横 187pixel にトリミングした．指標 N の相関値が向上したが，指標 A の相関値は下がってしまった．運転者の視線がどの程度偏るかは，人間の主観的要素であり，定量的に表すことが難しい．それに付随する課題として，相関値の改善も挙げられる．一般に 0.7 を超えた場合，強い相関関係にあるとされるため，手法の改善が望まれる．運転の難しさは，人間の主観による要因も大きく，熟練度にも大きく左右される．そのため，熟練度を分けた状態で研究を行うことで，相関値の向上が見込まれる．本研究では，一定範囲のみの可視化となったが，可視化の範囲を広げることで，運転が容易な地域の発見や運転経路の探索に役立てられる可能性がある．

謝辞

本研究は，JSPS 科研費 20K12081，野村マネジメントスクール研究助成及び東京都立大学傾斜的研究費 (全学分) 学長裁量枠国際研究環支援による．

参考文献

[1] 財団法人交通事故総合分析センター，“人はどんなミスをして交通事故を起こすのか”，イタルダインフォメーション，No.33, 2001

[2] Skurowski, Przemyslaw, and Marcin Paszkuta.: Saliency map based analysis for prediction of car driving difficulty in Google street view scenes. AIP Conference Proceedings. Vol.1978. No. 1. p. 110003, 2018

[3] Chi Zhang, Yuehu Liu, Qilin Zhang, et al.: A graded offline evaluation framework for intelligent vehicle ’s cognitive ability. IEEE Intelligent Vehicles Symposium (IV). pp. 320-325, 2018.

[4] Yuji Takeda, Kazuya Inoue, Motohiro Kimura, et al.: Electrophysiological assessment of driving pleasure and difficulty using a task-irrelevant probe technique. Biological Psychology Volume 120 pp. 137-141, 2016

[5] Cian Ryan, Finbarr Murphy, Martin Mullins: Spatial risk modelling of behavioural hotspots: Risk-aware path planning for autonomous vehicles. Transportation Research Part A 134, pp. 152–163, 2020

[6] Jianqiang Wanga, Yang Zhenga, Xiaofei Lia, et al.: Driving risk assessment using near-crash database through data mining of tree-based model. Accident Analysis and Prevention 84, pp. 54–64, 2015

[7] Frintrop, Simone, Erich Rome,, et al.: Computational visual attention systems and their cognitive foundations: A survey. ACM Transactions on Applied Perception (TAP) 7.1 pp. 1-39. 2010

[8] 三浦利章: 運転場面における視覚的行動 - 眼球運動の測定による接近 -, 大阪大学人間科学部紀要, Vol.3, pp.253-289. 1979

[9] Nakayasu, H., Seya, Y., Miyoshi, T., Keren, et al.: Measurement of visual attention and useful field of view during driving tasks using a driving simulator. Proceedings of the 2007 Mid-Continent Transportation Research Symposium. 2007

[10] Miltenburg, P. and Kuiken M.: The effect of driving experience on visual search strategies: Results of a laboratory experiment, University of Groningen, Haren. 1990

[11] Koch, C. and Ullman, S. : Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology 4, 4,pp. 219–227. 1985

[12] Hou, Xiaodi, and Liqing Zhang.: Saliency detection: A spectral residual approach. IEEE Conference on computer vision and pattern recognition. 2007

[13] Guo, Chenlei, Qi Ma, and Liming Zhang.: Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008.

[14] Schauerte, Boris, and Rainer Stiefelhagen.: Quaternion-based spectral saliency detection for eye fixation prediction. European Conference on Computer Vision. 2012.

ARG WI2 No.16, 2020
WI2-2020-02

SNSを利用した観光情報の特徴量化と地域間の類似性の分析

劉 瀚聡[†,a]　荒木 徹也[††,b]　遠藤 雅樹[†††,c]　石川 博[††††,d]

† 東京都立大学大学院システムデザイン研究科情報科学域　†† 群馬大学理工学府
††† 職業能力開発総合大学校基盤ものづくり系　†††† 東京都立大学大学院システムデザイン研究科

a) *liu-hancong@ed.tmu.ac.jp*　b) *tetsuya.araki@gunma-u.ac.jp*　c) *endou@uitec.ac.jp*　d) *ishikawa-hiroshi@tmu.ac.jp*

概要　SNS への投稿の中には，投稿者が実際に行った地域の写真や感想など，有益な情報が存在する．しかし，SNS には膨大な量の投稿がある．そのため，SNS を利用した自分好みの観光地の発見や類似性のある観光地同士の比較などが難しい．SNS を観光への分析に生かすには，観光に関する適切な情報を統合することが必要である．本研究では SNS 上の写真や投稿に添付されているタグ，位置情報などの情報を利用し，観光地ごとに特徴量を抽出する．また，それらを利用し，地域間の類似度を分析する．具体的には，Word2Vec を使用して地域ごとにタグの特徴量を抽出し，特徴量ベクトルを生成する．それにより，地域間の類似度の比較をすることができる．また，VGG16 を使用して投稿写真を処理し，類似度の高い観光地を発見する手法を提案する．

キーワード　Flickr，タグ，観光情報

1　はじめに

近年，SNS の普及により，迅速かつ容易に観光に関する大量な情報を手に入れることができるようになった．その中には，観光客が実際に訪問した観光地に対する感想や，観光地の撮影写真，評価，ハッシュタグ，位置情報など，投稿には様々な有益な情報が存在している．これらの過去の投稿はこれから旅行に出かける人の行動に影響を与える可能性がある．しかし，SNS には，膨大な情報が存在する．SNS を観光の参考にしている観光者にとっては，自分好みの目的地・観光地であるか，SNS を用いて判断することは一般に難しいと考えられる．SNS の中でも写真の共有を目的としたコミュニティウェブサイトである Flickr[1]が注目されている．Flickr では誰かがアップした写真に自由に「タグ」と呼ばれるキーワードを付けることができる．これらを分類することで，他のユーザーとタグを通したコミュニティを形成することができる．タグを用いることで，観光客が撮った旅行写真を Flickr 上で整理・分類・展示するほか，見知らぬ人と共有して互いにコメントを書き込むこともできる．

本研究ではタグを利用して，観光地の特徴の数値化を行う．また，観光地間の特徴量を比較することで類似性のある観光地を発見する手法を提案する．

本論文の構成は次の通りである．２章では本研究の関連研究を述べる．３章では，本研究の提案手法について述べる．４章では，実験結果を示す．５章では，本論文のまとめと今後の展望について述べる．

2　関連研究

本章では，関連研究について述べる．観光領域において SNS 上のタグや画像などのデータから，特徴量などの有益な情報を抽出する研究が多く行われている．

上原ら [1] は Web 上に混在する観光情報を活用し，多種類の情報から複数の特徴ベクトルを生成して観光地間の類似性を評価することで，観光地を推薦するシステムを提案した．酒井ら [2] は Flickr の投稿に基づき，観光客の嗜好の多様性を考慮し，その観光者が過去に撮影した写真の地域情報も組み込んだ協調フィルタリングを行い，その観光スポットを推薦するとともに，次に訪れる観光スポットを推薦するシステムの構築を実現した．

北村ら [3] は写真に取り付けられたタグの単語から撮影された位置の「魅力」を分析した．小原ら [4] は，ジオタグを利用して行動分析や観光情報の抽出の研究を行なった．長谷川ら [5] は，Twitter の投稿（以下ツイート）から地名との共起を利用して観光スポットの特徴語を抽出し，隣接スポットの特徴語との類似に基づいて観光地域を特定し，その特徴語辞書の構築手法を提案した．鈴木ら [6] は，Word2Vec を用いてツイートを学習し，教師あり学習を利用して道に迷っているときに発信されたと考えられる迷子ツイートの抽出を行い，可視化し，分析を行なった．

また，Yanai ら [7] は CNN に基づいた物体認識システムの効率的なモバイル実装を行なった．本研究は Word2Vec [8] モデルと CNN の一種である VGG16[9] を利用して Flickr に投稿されたタグと写真の特徴を抽出し，観光地間の類似性の分析を行なう．

Copyright is held by the author(s).
The article has been published without reviewing.

[1] https://www.flickr.com

3　提案手法

本章では，本研究で使用した観光情報の特徴量化手法と地域間の類似性を比較する方法について述べる．提案手法の概略を図 1 に示す．

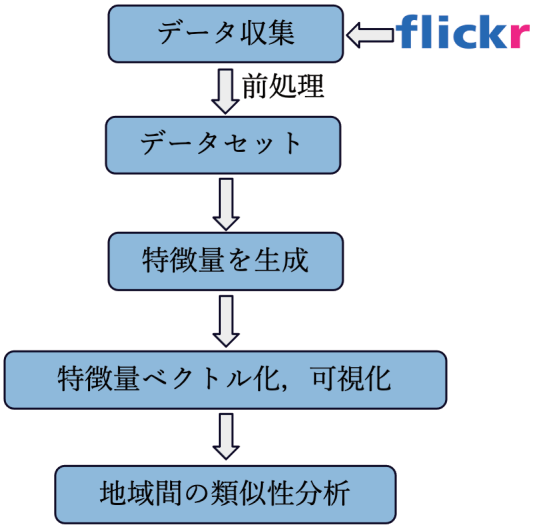


図 1 提案手法の概略

3.1　Flickr 投稿の収集と前処理

まず，2017 年分の Flickr 上の撮影写真，タグ，位置情報を収集し，緯度経度に基づいて地域ごとの投稿を抽出する．しかし，研究には写真投稿地の詳細住所が必要であるため，逆ジオコーディングを行う．逆ジオコーディングは，Nominatim[2]で OpenStreetMap[3]から 2017 年全部の投稿写真の緯度経度から詳しい住所を取得し，都道府県別で分類するという処理である．逆ジオコーディングの例を図 2 に示す．分類した後の写真の枚数を，図 3 に示す．

緯度経度	詳しい住所
[35.69664, 139.770578].	昭和シェル, 575, 神田須田町1, 神田須田町, 東京, 千代田区,東京都, 101-0033, 日本

図 2 逆ジオの例

3.2　特定の観光地のデータを抽出する

本研究は特定の観光地を対象にして分析を行うため，小さい地域内の Flickr の投稿を抽出する必要がある．逆ジオコーディングの結果に基づき，都道府県別で Flickr の投稿データセットが整理されたが，番地以降の住所は欠損値が多く，特定な範囲内の投稿を絞り込むのが難しいという問題がある．本研究では逆ジオコーディングの結果にある郵便番号を活用し特定の観光地の範囲を指定する．例を図 4 に示す．

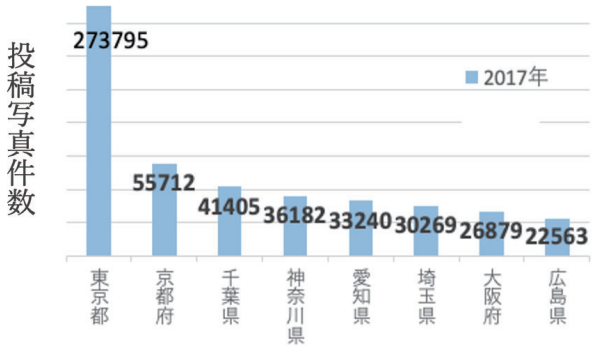


図 3 2017 年都道府県別の写真数

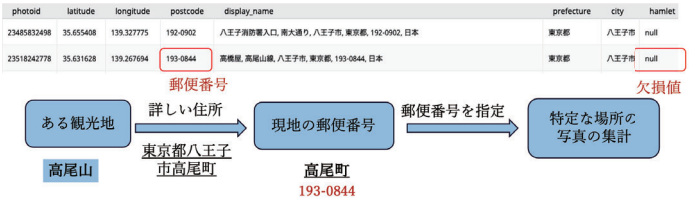


図 4 特定な観光地の写真を取得する流れ

3.3　特徴量の生成・可視化

本研究では TF-IDF を使用してタグのテキストに出現する単語の重要性を評価する．TF-IDF は，文書中に含まれる単語の重要度を評価する手法の 1 つであり，主に情報検索やトピック分析などの分野で用いられている． TF-IDF は，TF（英: Term Frequency, 単語の出現頻度）と IDF（英: Inverse Document Frequency, 逆文書頻度）の二つの指標に基づいて計算される．ある観光地に関する投稿内の単語の TF-IDF の値を算出して並べ替えを行い，上位 5 件の単語をこの観光地の特徴語とする．その後，それらの単語と TF-IDF 値を使用して，観光地の特徴レーダーチャートを生成する．

3.4　地域間の類似性の分析

本研究では自然言語処理ツールである Word2Vec と CNN（Convolutional Neural Network）の代表的なモデルである VGG16 を導入し，それぞれ文書と画像を分析して地域間の類似性を算出する．これにより，類似性の高い観光地の発見が期待できる．

3.4.1　テキスト処理のモデルで類似性を分析

本節では Word2Vec モデルを導入して Flickr の投稿写真に付けられたタグの分析について述べる．

Word2Vec は単語をベクトル化する手法で，共起回数

[2]https://wiki.openstreetmap.org/wiki/JA:Nominatim
[3]https://openstreetmap.jp/.

を扱うのではなく，単語の周辺に出現する単語の確率を扱うという自然言語処理のツールである．本研究では Word2Vec で 2017 年分の Flickr の投稿写真に付けられていたタグの集計を訓練し，コーパスを用意する．これにより，ある単語と類似している単語を見ることができ，２つの単語の類似度を比較することができる．例えば，「上野公園」,「井の頭公園」との単語を入力すると，この２つの単語の類似度を比較することができ，この２つの観光地の類似度を算出できる．

3.4.2　CNN モデルで類似性を分析

Deep Learning 分類機の一種である CNN は画像認識の課題において汎用性が高く，特徴抽出機として他の用途に活用もできる．そこで，本研究では代表的な CNN モデルである VGG16 を導入する．

Flickr の投稿写真の特徴量を抽出して特徴ベクトルを生成し，特徴ベクトル間のコサイン類似度 [1] を算出する．２つの画像の特徴ベクトル $\vec{a} = (a_1, a_2, \cdots, a_n)$ と，ベクトル $\vec{b} = (b_1, b_2, \cdots, b_n)$ のコサイン類似度 R_{AB} は次のように表される．

$$R_{AB} = \frac{\sum_{i=0}^{n} a_i b_i}{\sqrt{\sum_{i=0}^{n} a_i^2}\sqrt{\sum_{i=0}^{n} b_i^2}} \quad (1)$$

このモデルを使用し，まず A 観光地（以下 A 地）の投稿写真のデータセットと，B 観光地（以下 B 地）の投稿写真のデータセットを用意する．そして A 地における like 数が一番多い写真を VGG16 モデルに入力し，B 地のデータセットからコサイン類似度が最も高い写真を出力させる．それにより，B 地から A 地との一番類似の高い写真を抽出できる．

4　実験

4.1　観光特徴量の生成・可視化

3.3 節で説明した方法を使用し，Flickr 上の 2017 年東京上野公園周辺の投稿写真を例にとり，合計 5088 枚を収集した．それらの写真に付けられていたタグを抽出し，MeCab で形態素解析を行なって単語に分割した．そして単語の TF-IDF 値を算出して並べ替え，TOP5 の特徴単語を抽出した．取得された特徴単語と TF-IDF 値は図 5 に示す．レーダーチャートを図 6 に示す．

特徴単語	zoo	park	sakura	不忍池	lotus
TF-IDF	0.2732	0.0940	0.0489	0.0477	0.0395

図 5 例：上野公園の観光特徴量

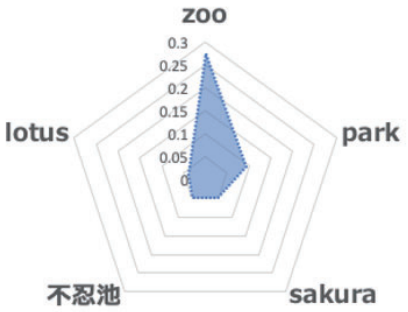


図 6 例：上野公園観光特徴のレーダーチャート

4.2　地域間の類似性の分析，類似観光地を発見

本研究では，東京タワーとの類似度が高いタワーのある観光地を発見するという実験を行なった．

まず，NAVITIME[4]から上位 10 件のタワーのある観光地とそれぞれの位置情報を収集し，その 10 件の Flickr の投稿写真のデータセットを作成した．観光地名と写真数を図 7 に示す．

次に，東京タワー周辺の like 数が最も高い写真を代表的な写真とした．そして 3.4.2 節で説明した VGG16 の特徴抽出機に東京タワーの代表的な写真を入力し，他のタワー観光地ごとに写真を比べ，コサイン類似度の上位３件の写真を抽出し，コサイン類似度の平均値をとって，そのタワーのある観光地が東京タワーの類似度算出の結果とした．湘南平（高麗山公園）展望台の例を図 8 に示す．その後，10 箇所のタワーのある観光地それぞれの類似度結果を元に並べ替え，順番が１位の「さっぽろタワー」が東京タワーと類似度の高い観光地であるという結果になった．実験結果を図 9 に示す．

この手法を活用し，別の類似な観光地を見つける課題も汎用できる．

観光地	写真数
さっぽろテレビ塔	265
名古屋テレビ塔	168
湘南平（高麗山公園）展望台	44
スカイツリー	5,873
神戸タワー	191
別府タワー	71
博多ポートタワー	122
福岡タワー	109
京都タワー	1,143
通天閣	1,286

図 7 タワー観光地データセット

[4]https://www.navitime.co.jp

TOP3	画像類似度
38613202902.jpg	0.734365
38803971981.jpg	0.721892
37918016435.jpg	0.703879
平均値	0.720045333

図 8 湘南平（高麗山公園）展望台の類似写真

順番	観光地	東京タワーとの類似度
①	さっぽろテレビ塔	0.758215000
②	名古屋テレビ塔	0.724012667
③	湘南平（高麗山公園）展望台	0.720045333
④	スカイツリー	0.712422333
⑤	神戸タワー	0.712133667
⑥	別府タワー	0.711953667
⑦	博多ポートタワー	0.708843667
⑧	福岡タワー	0.690292667
⑨	京都タワー	0.673201333
⑩	通天閣	0.641043000

図 9 タワー観光地の類似度ランキング

4.3 統合した手法で観光地間の類似性の評価

本研究では Flickr の 2017 年全部の投稿に付けられたタグのテキストを，3.4.1 節で説明した Word2Vec モデルで訓練してコーパスを生成した．そして「東京タワー」と「さっぽろタワー」の単語を例にとって入力し，類似度を取得した．次に，VGG16 モデルの結果と Word2Vec モデルの結果を統合し，画像認識分野と自然言語処理分野両方を考えた手法で地域間の類似度を分析した．例を図 10 に示す．

観光地	VGG16モデル類似度	Word2Vec類似度	平均類似度
東京タワー＆さっぽろテレビタワー	0.7582150	0.6610698	0.7096424

図 10 統合した類似度

5 まとめと今後の課題

本研究は Flickr に投稿された位置情報，タグ，写真などの情報を用いて，写真の緯度経度から逆ジオコーディングを行なって詳細な住所を取得した．特定の観光地のタグを利用し，TF-IDF を使用してタグに出現する単語の重要性を評価した．また，Word2Vec のモデルを利用してタグを訓練し，地域間の類似性を評価した．代表的な CNN モデルである VGG16 を導入し，画像のベクトルを生成し，コサイン類似度 R_{AB} を算出することで，特定の地域の類似度が高い写真を抽出した．最後に東京タワーとの類似度が高いタワーのある観光地を発見するという実験を行い，類似する観光スポットを見つける手法を提案した．

今後の課題として，より多種類のデータと別の SNS に対して，総合的な手法で観光特徴量を生成することが挙げられる．

謝辞

本研究の一部は，JSPS 科研費 20K12081，野村マネジメントスクール研究助成及び東京都立大学傾斜的研究費 (全学分) 学長裁量枠国際研究環支援による．

参考文献

[1] 上原尚，嶋田和孝，遠藤勉：Web 上に混在する観光情報を活用した観光地推薦システム，電子情報通信学会技術研究報告，言語理解とコミュニケーション研究会 (NLC)，第 4 回集合 知シンポジウム，NLC2012-35, pp.13-18, 2012.

[2] 酒井勇人，熊野雅仁，木村昌弘:Flickr データに基づいたインタラクティブ観光スポット推薦システム，SIG-AM, 14, 05, pp.24-29, 2016.

[3] 北村武士，本間健太郎，今井公太郎：Flickr のジオタグ付き写真データからみる日本全土の観光特性　居住国推定とタグクラスタリングによる訪日外国人の興味分析，日本建築学会計画系論文集，84, 755, pp.187-197, 2019.

[4] 小原基季，森田和宏，泓田正雄，ほか：Twitter 本文を用いた観光情報抽出及び分析システムの構築，人工知能学会全国大会論文集，29, pp.1-3, 2015.

[5] 長谷川馨亮，馬強，吉川正俊：Twitter からの地域特徴語辞書の構築とその観光情報検索への応用，第 6 回データ工学と情報マネジメントに関するフォーラム，2014.

[6] 鈴木亮平，廣田雅春，荒木徹也，ほか：位置情報付きツイートを用いた観光地周辺の迷いやすいスポットの発見，研究報告データベースシステム（DBS）2019-DBS-169,2,pp.1-6, 2019.

[7] Keiji Yanai, Ryosuke Tanno, and Koichi Okamoto. : Efficient mobile implementation of a cnn-based object recognition system. In Proceedings of the 24th ACM International Conference on Multimedia, MM ’ 16, pp.362–366, 2016.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean.: Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pp. 3111–3119, 2013.

[9] K. Simonyan and A. Zisserman.: Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.

ARG WI2 No.16, 2020
WI2-2020-03

大規模ソーシャルデータを用いた寄り道ルート推薦手法

夏 思浩　　井嶋 蒼　　横山 昌平
東京都立大学大学院システムデザイン研究科
xia-sihao@ed.tmu.ac.jp ijima-so@ed.tmu.ac.jp shohei@tmu.ac.jp

概要　LBSN(Location Based Social Network) の発展により，情報量が急激に増加し，ユーザが自ら POI(Point of Interest) をフィルタイングすることが困難である．そこで，ユーザーに適した場所を判別して提示するルート推薦が求められる．本論文では大規模 Flickr データを用い，ユーザが指定した出発地と目的地より，ある範囲内のユーザのルート情報を抽出し，ルート行列と有向グラフを生成し，ワーシャル-フロイド法を使い，ユーザが与えた二点間に存在する適した寄り道ルート順位 5 位を発見する．そして，発見した寄り道スポットを Google Places API で Venue を取得し，その中から美食ルートと買い物ルートを推薦する．

キーワード　LBSN，Flickr，寄り道ルート推薦，Warshall–Floyd Algorithm

1　はじめに

近年，スマートフォンの普及により、人々は日常生活や旅行先などにおいて，興味を持つものを気軽に撮影できるようになった．それらの写真には撮影した場所の緯度経度や時間などの情報が付いている．また、Flcikr や Foursquare などの LBSN(Location Based Social Network) の発展により、情報量が急激に増加し、ユーザが自ら POI(Point of Interest) をフィルタイングすることが困難になった。そこで、ユーザーに適した場所を判別して提示するルート推薦が求められる。

そこで本研究では，大規模 Flickr データに基づいて，ユーザが与えた二点間に存在する寄り道に適したスポットを推薦する．寄り道に適したスポットとは，ユーザが行く可能の場所を指す．具体的には，DBSCAN によりスポットを分け，出発地と目的地の距離を半径として，二つの円が重なる部分を候補範囲とする．その範囲にあるスポットの間の移動を Flickr データから抽出し，有向グラフに生成する．有向グラフの重みは頂点（スポット）間の移動数の逆数とする．最後はワーシャル-フロイド法により，出発地と目的地の間の最短経路順位 5 位までを発見する．そして，発見した寄り道スポットを efficient-geo-crawler[4][1]で Venue を取得し，その中から美食ルートと買い物ルートを推薦する．

本論文の構成は次の通りである．2 章では本研究の関連研究について述べる．3 章では提案手法について述べる．4 章では提案手法の実行例を示し，それに対する評価を述べる．5 章では本研究で得られた成果をまとめる．

Copyright is held by the author(s).
The article has been published without reviewing.

[1]https://github.com/YokoyamaLab/efficient-geo-crawler

2　関連研究

石倉ら [1] は，一般的なユーザからの総合的な評価を元にした店舗に対する数値，現在地から経由地を挟んだ目的地までの距離と経由地を挟まない直行距離の差分，道のりに含まれる信号機により受ける障害を計算した数値を求め，これらの値より最終的な推薦値を算出し，推薦経路を提案した．青山ら [2] は，ソーシャルメディアサイトに投稿された写真のジオタグと撮影日時を用いて，場所に対する人々の知名度と興味の度合いを算出し，それらに基づいて出発地から目的地までの移動中に立ち寄ることが可能な場所を発見する．ユーザが指定した出発地と目的地よりも知名度が低い場所の中から，興味の度合いが最も高い場所を寄り道候補として発見し，ランキング形式で提示するシステムを構築した．これらの研究においても，推薦値からある寄り道スポットを推薦する．本研究では，ユーザの移動数が多いルートから寄り道スポットを推薦する．

酒井ら [3] は，ユーザーが観光地域を指定したとき，ヒトの嗜好の多様性を考慮して，そのユーザーが過去に撮影した写真の地域情報も組み込んだ協調フィルタリングにより，その観光スポットを推薦するとともに，次に訪れる観光スポットを順次推薦するシステムを構築した．この研究においても，そのユーザの過去のデータから次に行く場所を推薦する．本研究では，全てのユーザの過去のデータからスポットを推薦する．

3　提案手法

本研究では，Flickr に投稿された写真の緯度経度や時間などのメタデータを用いて，DBSCAN によりスポットを分ける．ユーザの移動が有向グラフに生成し，その有向グラフの重みがユーザの移動数の逆数とする．ワーシャル-フロイド法により，現在地と目的地の間の最短経路順位 5 位までを推薦する．

3.1 データの収集

本研究では，Flickr に投稿された写真の緯度経度や時間などのメタデータを用いるため，Flickr API を用いて，2016 年から 2019 年までの 4 年間のデータを収集する．収集範囲は新宿を中心とする半径 32km の部分である．この API には，一度に大量の画像を検索しようとする場合に，途中から重複する画像が出てきてしまうという仕様がある. この問題に対して，今回は例えば「2016 年 1 月 1 日 0 時 0 分 0 秒〜2016 年 1 月 1 日 23 時 59 分 59 秒」のように 1 日ずつ区切り，データを収集した．取集したデータ数は重複を除いて 386,122 件である．

3.2 スポットの生成と凸包

本研究では DBSCAN を使い，収集した Flickr データをクラスタリングし，スポットを分ける．eps を 0.0001 ずつ変化させたクラスタリングの結果を表 1 に示す．DBSCAN のパラメータ eps を 0.0002，minsamples が 10 にした時に，クラスターの数が一番平均だったため，このパラメータを設定した．

minsamples = 10					
eps	0.0001	0.0002	0.0003	0.0004	0.0005
クラスター					
できた数	186980	263635	296833	314880	326613
できない数	189142	112487	79289	61242	49509
クラスターの数	3154	2888	2461	2129	1863
3 万以上の数	0	0	0	0	2
1 万〜3 万の数	1	4	6	6	2
1 千〜1 万の数	17	30	29	28	27

表 1 DBSCAN 対照

候補確定とスポットの可視化を行うため，そのスポットのバウンドが必要である．そこでグラハム探索を用いて，スポットの凸包とそのバウンドの頂点の緯度経度を探す．凸包とは，点集合 Q の各点がその境界上にあるような最小の凸多角形 P のことを指す．点集合 Q=p0,p1,...,p11 と凸包 CH(Q) を図 1 に示す．この例の凸包の頂点は p0,p4,p11,p9,p7 である．凸包を計算するアルゴリズムはいくつが存在するが，今回はグラハム探索 (Graham’s scan) を使って，凸包とそのバウンドの頂点の緯度経度を探す.

3.3 範囲内のルート情報の抽出

同一のユーザが同一のスポット間の移動を抽出する．ここで，例えば A → B → C などの移動は A → B と B → C がもちろん，A → C も抽出した．抽出したルート数の合計は約 22 万である．範囲には，ユーザが指定した出発地と目的地の距離を半径 R として，二つの円が重なる部分を候補範囲とする．3.2 章から計算したスポットの各頂点に関して，出発地と目的地それぞれの距離 d1,d2 を計算し，d1,d2 が R より低い頂点が存在したら，このスポットが候補になる．次に，22 万のルートから，候補範囲内のルートを抽出する．

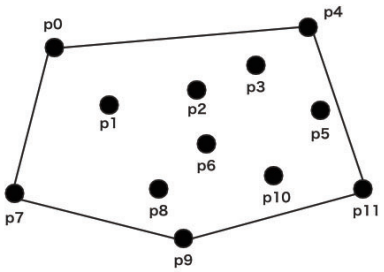


図 1 点集合 Q とその凸包 CH(Q) の例

3.4 ルート行列と有向グラフの生成

抽出したルートデータを用いて二次元行列と有向グラフを生成する．その有向グラフの重みは行列に対応数字の逆数とする．有向グラフの例を図 3 に示す．ポリゴンはスポットを表しており，有向グラフに対しては頂点になる．頂点間の重みは実際移動数の逆数になる．

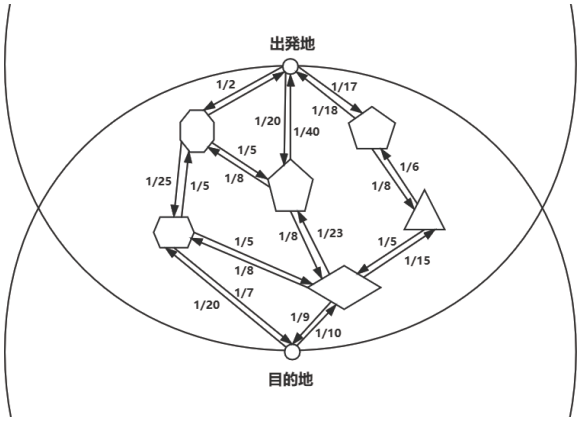


図 2 有向グラフの例

3.5 ワーシャル-フロイド法で最短経路

ワーシャル-フロイド法とは，重み付き有向グラフの全ペアの最短経路問題を多項式時間で解くアルゴリズムである．グラフの頂点を $V = 1, ..., n$ とし，n × n 配列 d の (i,j) 成分をグラフの i,j 間の枝長 (枝がなければ ∞, i=j はゼロ) で初期化してから以下の 3 重ループを実行すると，すべての頂点 i,j についてそれらの間の最短路長が d[i,j] に入る．

4 実行例と考察

提案手法の実行結果を示し，結果の考察を行う．今回は上野駅と浅草駅それぞれが出発地と目的地となり，この間の寄り道候補スポットとルートの抽出を行う．抽出するデータはワーシャル-フロイド法で計算し，最短順位 5 位までのルートを推薦する．この五つのルートから経由したスポットを efficient-geo-crawler で Venue を取

得し，成分を考察する．最終的な結果は leaflet で可視化する．

4.1 実行結果

上野駅と浅草駅の候補スポットの可視化を図 4 に示す．青色のマーカーは上野駅と浅草駅を表している．赤色のポリゴンは上野駅と浅草駅の間の候補スポットを表している．候補スポットの数は 92 個が存在し，上野駅が所属するスポットと浅草駅が所属するスポットを加えると，総合 94 個のスポットから 94×94 の二次元行列と有向グラフを生成する．ただし，上野駅から浅草駅までのルートと浅草駅から上野駅までのルートは 0 とする．

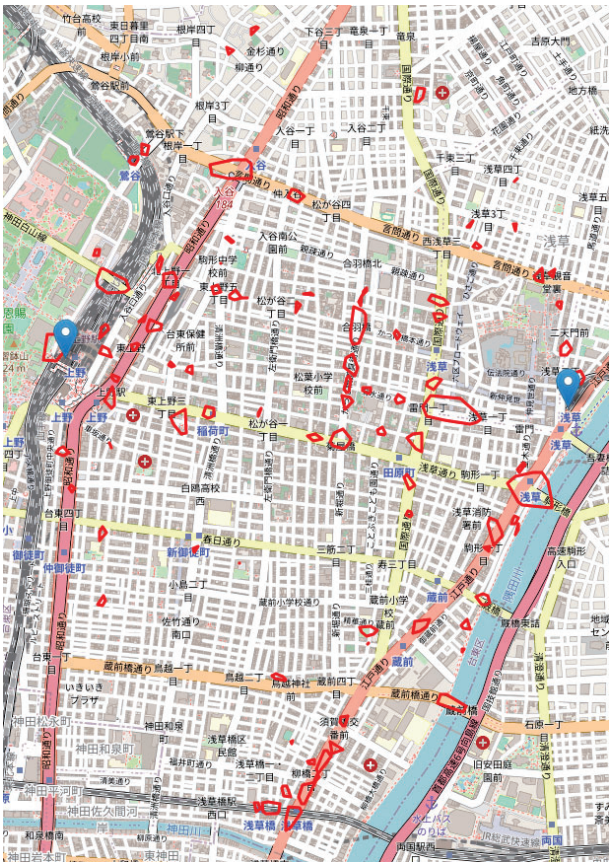


図 3 候補スポット

4.1.1 浅草駅→上野駅

まずは浅草駅から各候補スポットまでの移動数と各候補スポットから上野駅までの移動数を二次元行列に埋め込む．次に，行列の 0 を無限大に変更し，0 ではない数字を逆数に変更する．ワーシャル-フロイド法により，任意の二つの頂点の最短経路を計算する．次に，浅草駅が所属するスポットから上野駅が所属するスポットまでの順位 5 位のルートを可視化する．順位 5 位までのルートの可視化を図 5 に示す．推薦された寄り道スポットは雷門一丁目 (スポット 415)、上野公園バス駐車場 (スポット 191)、菊屋橋 (スポット 517)、合羽橋道具街 (スポット 360)、駒形橋西詰 (スポット 516) である．最後は efficient-geo-crawler を使って Venue の総数や restaurant と store の数を取得する．結果を表 2 と図 6 に示す．

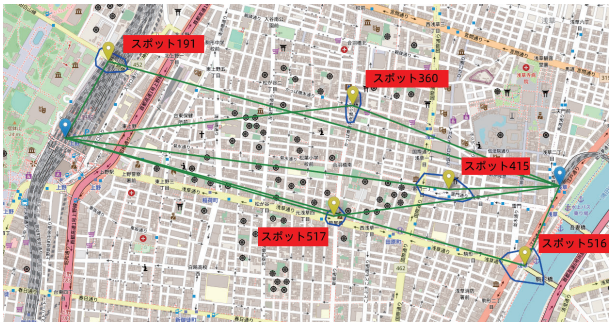


図 4 浅草駅→上野駅

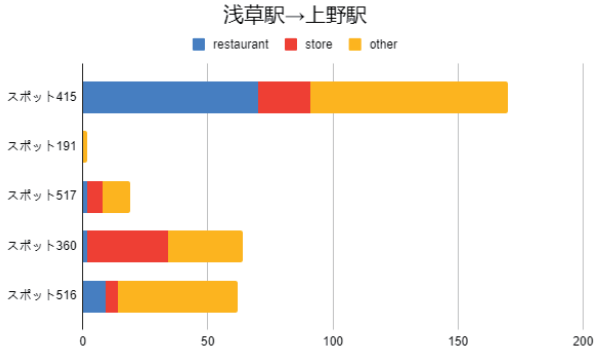


図 5 各スポットの Venue

推薦順位	1 位	2 位	3 位	4 位	5 位
推薦されたスポット	415	191	517	360	516
浅草駅からの移動数	73	8	14	12	50
上野駅までの移動数	8	71	12	10	6
全部の Venue	170	2	19	64	48
restaurant	70	0	2	2	9
store	21	0	6	32	5

表 2 浅草駅→上野駅

4.1.2 上野駅→浅草駅

4.1.1 章と同じような，順位 5 位までのルートの可視化を図 7 に示す．推薦された寄り道スポットは雷門一丁目 (スポット 415)、上野駅前 (スポット 294)、上野公園バス駐車場 (スポット 191)、駒形橋西詰 (スポット 516)、合羽橋道具街 (スポット 360) である．efficient-geo-crawler で取得する結果は表 3 と図 8 を示す．

4.2 考察

図 4 および表 2 により，2 位の上野公園バス駐車場 (スポット 191) は Venue が 2 件のみであり，restaurant と store に関しても 0 件である．これは旅行客が乗降する場所であると推測できるため，推薦ルートから除外する．また，1 位の雷門一丁目 (スポット 415) の restaurant の数が一番多い，全体 Venue に占める比率も一番多いの

推薦順位	1 位	2 位	3 位	4 位	5 位
推薦されたスポット	415	294	191	516	360
上野駅からの移動数	13	43	58	6	7
浅草駅までの移動数	78	14	9	56	19
全部の Venue	170	6	2	48	64
restaurant	70	0	0	9	2
store	21	0	0	5	32

表 3 上野駅→浅草駅

で，このルートは美食ルートとして推薦するのが良いと考えられる．また，4 位の合羽橋道具街 (スポット 360) の store の数が一番多く，全体 Venue の比率も一番多いので，このルートは買い物ルートとして推薦するのがいいと考えられる．図 6 および表 3 より，美食ルートはスポット 415，買い物ルートはスポット 360 が推薦されている．両方の結果より，ほぼ同じような寄り道スポットが出ているため，雷門一丁目 (スポット 415) と合羽橋道具街 (スポット 360) は浅草駅と上野駅の間の適した寄り道スポットと考えられる．

表 2 より，浅草駅から各スポットまでの移動数と各スポットから上野駅までの移動数は差がある．これは駅との距離に正比例すると考えられる．また，3 位と 4 位の両方の移動数の和は 5 位より少ないが，順位は高い．これは逆数が原因である．本研究では両方の移動数が多いと適した寄り道ルートと設定するので，5 位のような片方だけが多いルートは両方が多いより適した寄り道ルートではない．

また，今回の実験結果では，全部の寄り道ルート候補は一つだけのスポットを含まれる．これは二つの原因が考えられる．一つ目はデータ量の少ないことである．移動数が増えれば，複数スポットを含まれるルートが出るかもしれない．二つ目は出発地と目的地が近いことである．例えば新宿駅と東京駅の場合，有向グラフの頂点が多くなり，複数スポットを通る可能性も高くなると考えられる．

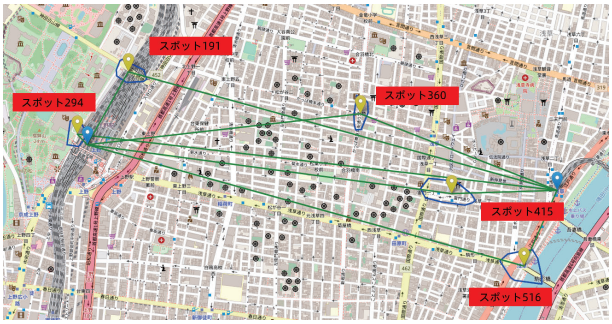


図 6 上野駅→浅草駅

5 おわりに

本研究では，大規模 Flickr データに基づいて，ユーザが与えた二点間に存在する寄り道に適したスポットの推薦手法を提案した．提案手法では，ユーザが指定した出発地と目的地より，ある範囲内のユーザのルート情報を抽出する．そして，ルート行列と有向グラフを生成し，ワーシャル-フロイド法を使い，適した寄り道ルート順位 5 位までを発見する．そして，発見した寄り道スポットを efficient-geo-crawler で Venue を取得し，その中から美食ルートと買い物ルートを推薦する．また，上野駅と浅草駅それぞれを出発地と目的地に指定して提案手法を実行し，結果に対して考察を行った．

今後の課題として，データをより多く収集し，時間帯を分けて推薦を行うことを考えている．

参考文献

[1] 石倉頌子，小林亜樹: 寄り道経路推薦方式，第 74 回全国大会講演論文集，2012, 1, pp. 297 - 298, 2012.

[2] 青山賢，廣田雅春，石川博，横山昌平: ジオタグ付き写真を用いた知名度が低いにもかかわらず興味の度合いが高い寄り道候補の発見，DEIM Forum 2015, B5-3, 2015.

[3] 酒井勇人，熊野雅仁，木村昌弘: Flickr データに基づいたインタラクティブ観光スポット推薦システム， インタラクティブ情報アクセスと可視化マイニング研究会, 14, 05, pp. 24 - 29, 2016.

[4] Ijima.S, Hirota.K, Yokoyama.S: A Crawling Method with No Parameters for Geo-social Data based on Road Maps, Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services, pp. 250-254, 2019.

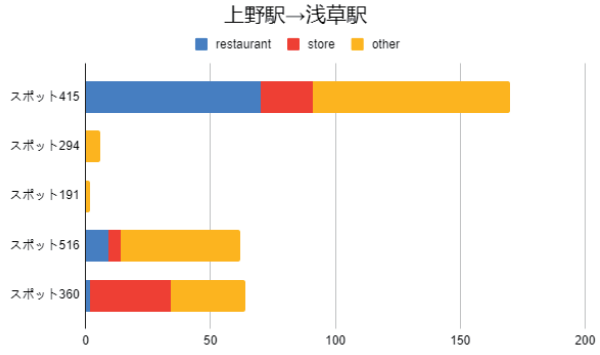


図 7 各スポットの Venue

ARG WI2 No.16, 2020
WI2-2020-04

散策時の発見性向上のための訪問経験に基づく動的散策マップシステムの検討

塩崎 イサム　　奥 健太
龍谷大学 理工学部
t170485@mail.ryukoku.ac.jp okukenta@rins.ryukoku.ac.jp

概要　地図ベースのインタフェースを採用した観光情報推薦システムは多いが，地図上に過度に情報を提示することにより，散策時の発見性を損なう問題もある．本研究では，散策時の発見性を促すために，ユーザの訪問経験に応じてマップを隠蔽/解放する動的散策マップシステムを提案する．本システムでは，初期状態はマップのすべての領域を隠蔽しておき，その後，ユーザの訪問経験に応じて，隠蔽されていた領域を解放していく．このような要素により，散策に対する娯楽性の向上とともに，散策時の発見性の向上が期待できる．

キーワード　観光情報推薦システム，発見性，散策マップシステム

1　はじめに

ドライブやサイクリング，トレッキング，ウォーキングなどを楽しむ人々は多い．このような人々の中には，その道程自体を楽しむことを主目的とすることも多い．さらには，その楽しみ方の一つとして，これまでの軌跡を記録することの楽しみ，新規ルートを開拓することの楽しみも挙げられる．特に，これまでに通ったことのないルートを通ることで，新たな景観に出会ったり，新たな店を発見したりという喜びにつながる．

地図ベースのインタフェースを採用した観光情報推薦システムは多い [1]．Lee らのシステム [2] は，Google マップ[1]を用いて，地図上に旅行ルートを提示する．他にも，地図ベース観光情報推薦システムとして，e-Tourism[3] や City Trip Planner[4]，Otium[5] などが提案されている．また，CT-Planner[6] はユーザの嗜好や要求に合った旅行プラン情報を地図上に提示する．ここでは，訪問すべきスポットやスポット間を結ぶルートなどがナビゲーション情報と共に地図上に詳細に提示される．観光情報には観光スポットや宿泊施設，飲食店，ルートなど，位置情報に密接に関連した情報が多く含まれるため，これらを地図上に提示することはユーザにとって直感的に理解しやすい．これらのシステムは特に見知らぬ地域に訪れたユーザにとっては有用である．しかしながら，過度に情報を提示することにより，ユーザにとって旅行時の発見性を損なう問題もある．

Izumi ら [7] は，旅行時の発見性を促すために，黒塗り地図システムを提案している．彼らのシステムでは，初期状態では地図のすべてを黒塗りしておき，ユーザが実際に通過した部分のみが表示される．したがって，ユーザは自身が通過した領域の情報のみしか得られず，未通過の領域の情報は確認できない．この要素により，未通過の領域に対してユーザの興味を誘発し，探索意欲を向上させようというのが彼らのシステムの目的である．

本研究でも Izumi らの動機を参考に，ユーザの訪問経験に応じてマップを隠蔽/解放する動的散策マップシステムを提案する．本システムでは，初期状態はマップのすべての領域を隠蔽しておき，その後，ユーザの訪問経験に応じて，隠蔽されていた領域を解放していく．このような要素により散策に対する娯楽性の向上とともに，散策時の発見性の向上が期待できる．本稿では，試作した動的散策マップシステムについて説明する．

2　定義

本稿で用いる用語の定義を以下に示す．また，本稿で用いる記号を表 1 にまとめる．

道路ネットワーク．　道路ネットワークは有向重み付きグラフ $G = (V, E)$ で表現される．ここで，V は道路ノード集合であり，$E \subseteq V \times V$ は道路リンク集合である．道路ノード $v_i \in V$ は交差点や道路の終端を表す．道路リンク $e_{i,j} = (v_i, v_j) \in E$ は，始点ノード v_i から終点ノード v_j へ向かう有向リンクである．

隣接行列．　道路ネットワーク G はノード間の隣接行列としても表すことができる．隣接行列を行列 $A = [a_{i,j}]^{|V|\times|V|}$ で表す．ここで，$a_{i,j} = 1$ のとき，二つのノード v_i, v_j は隣接するといい，$e_{i,j} = (v_i, v_j)$ が存在する．

メッシュ．　対象エリアを緯度・経度に基づいて同一の大きさでメッシュ状に分割したものをメッシュとよぶ．対象エリアを分割したメッシュ集合を $\mathcal{M}$ と表し，メッシュ $k \in \mathcal{M}$ の矩形領域を M_k とする．

Copyright is held by the author(s).
The article has been published without reviewing.

[1] https://www.google.co.jp/maps/?hl=ja

表 1 本稿で用いる記号.

記号	説明
$G=(V,E)$	道路ネットワーク.
$i \in \{1,\ldots,\mid V \mid\}$, $j \in \{1,\ldots,\mid V \mid\}$	道路ノードのインデックス.
$v_i \in V$	道路ノード.
$e_{i,j}=(v_i,v_j) \in E$	道路リンク.
$A=[a_{i,j}]^{\mid V \mid \times \mid V \mid}$	隣接行列.
$\mathcal{M}$	メッシュ集合.
$k \in \{1,\ldots,\mid \mathcal{M} \mid\}$	メッシュのインデックス.
$M_k \in \mathcal{M}$	メッシュk の矩形領域.
$c \in \{1,\ldots,\mid V \mid\}$	現在地ノードのインデックス.
N	スポット数.
$l \in \{1,\ldots,N\}$	スポットのインデックス.
$s_i \in \{0,1,2\}$	ノード i の状態ラベル.
$t_{i,j} \in \{0,1\}$	リンク $e_{i,j}$ の状態ラベル.
$u_l \in \{0,1,2\}$	スポット l の状態ラベル.
$m_k \in \{0,1\}$	メッシュM_k の状態ラベル.

M_k は一辺の長さ d の矩形領域で表され，その南西緯度・経度と北東緯度・経度を属性にもつ.

ノードの状態ラベル. ノード i の状態ラベルを $s_i \in \{0,1,2\}$ で表す．$s_i=0$ のときノード i を未訪問，$s_i=1$ のとき訪問済みであることを表し，$s_i=2$ のときノード i が訪問済みノードに隣接するノードであることを表す.

リンクの状態ラベル. リンク $e_{i,j}$ の状態ラベルを $t_{i,j} \in \{0,1\}$ で表す．$t_{i,j}=0$ のときリンク $e_{i,j}$ を未通過，$t_{i,j}=1$ のとき通過済みであることを表す.

スポットの状態ラベル. スポット l の状態ラベルを $u_l \in \{0,1,2\}$ で表す．$u_l=0$ のときスポット l は非公開，$u_l=1$ のとき公開，$u_l=2$ のとき訪問済みであることを表す.

メッシュの状態ラベル. メッシュk の状態ラベルを $m_k \in \{0,1\}$ で表す．$m_k=0$ のときメッシュk を隠蔽状態，$m_k=1$ のとき解放状態であることを表す.

3 動的散策マップシステムの概要

本章では，提案システムである動的散策マップシステムの概要を説明する．図 1 はシステムのインタフェースである．本システムでは，初期状態はマップのすべての領域を隠蔽しておき，その後，ユーザの訪問状態に応じて，隠蔽されていた領域を解放していく．また，システムには「おうちで」モードと「リアル散策」モードの二つのモードを用意している．本稿では,「おうちで」モードについて説明する.

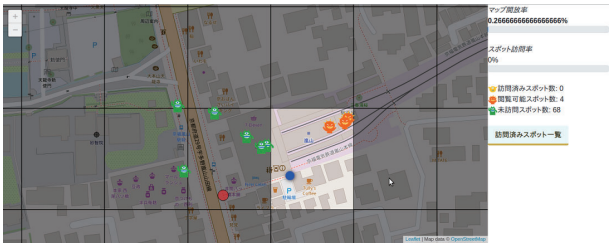


図 1 システムのインタフェース．地図画像は Leaflet API を用いて OpenStreetMap の画像を取得した.

ユーザは自宅で PC やタブレット端末などのデバイスを用いて本システムを利用する．システムには京都御所や嵐山といった小規模の散策エリアがあらかじめ登録されている．ユーザが散策したいエリアを選択すると，そのエリアマップがメインに表示される．ただし，初期状態では詳細な情報は表示されない．初期状態では，任意のいくつかのノードが開始地点候補として表示されている．ユーザはこの中から一つのノードを開始地点（以降，開始ノードとよぶ）として選択する．すると，選択されたノードが現在地として強調表示される.

ユーザはこの開始ノードから散策を進めていく．初期状態では，図 2(a) のように開始ノード付近を除き，すべての領域が隠蔽されている．マップ上には開始ノードに隣接するノードが表示されている．ユーザは表示されているノードのうちいずれか一つをクリック（またはタップ）することで擬似的に移動する．すると，選択されたノードが新たな現在地として強調表示される．同時に，図 2(b) のように，新たな現在地周辺の領域が解放される．領域の解放とともに，その解放された領域に含まれるスポット（店舗や観光スポットなど）も表示される.

このような操作を繰り返していくことで，ユーザは擬似的なエリア散策を楽しむことができる．図 2 のように，散策を進めていくことで，動的に隠蔽されていた領域が解放されていく．このような要素により，ユーザがこれまでに通ったことのなかったルートやその沿道にあるスポットに気付き，実際に現地を訪れてみたくなるような興味の誘発につながることが期待できる.

4 ノードの訪問状態に基づくマップの隠蔽/解放

4.1 対象エリアの道路ネットワークデータの構築

本システムは京都御所や嵐山といった小規模の散策エリアを対象としている．ここでは，便宜上，対象とする散策エリアの道路ネットワークを $G=(V,E)$ とする．対象エリアを含む領域をポリゴンとして作成し，全道路ネットワークからそのポリゴン領域に含まれる部分道路ネットワークを切り出し，それを G とする．図 3 に嵐山の道路ネットワークの例を示す．図中の点はノード

(a) 初期状態.　　(b) 隠蔽領域の解放.

図 2　初期状態は (a) のように，スタート地点付近を除きマップのすべての領域が隠蔽されている．ユーザが散策を進めていくと，(b) のように訪問経験に応じて動的に隠蔽されていた領域が解放されていく．地図画像は Leaflet API を用いて OpenStreetMap の画像を取得した．

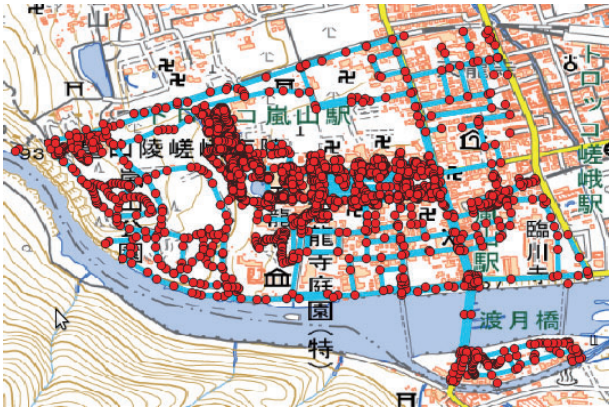


図 3　嵐山の道路ネットワーク．点はノード $v_i \in V$ を，線はリンク $e_k \in E$ を表す．地図画像は OpenStreetMap の画像をキャプチャした．

$v_i \in V$ を，線はリンク $e_{i,j} \in E$ を表す．

4.2　ノードおよびリンクの訪問状態の更新

道路ネットワーク G において，ユーザがノード i を訪問すると，これに応じて各道路ノードの状態 s_i およびリンクの状態 $t_{i,j}$ を更新する．アルゴリズム 1 に更新アルゴリズムを示す．

まず，o には開始ノードのインデックスを入力する．3 行目から 10 行目は初期化処理である．現在地ノード c は NULL で初期化しておく．つづいて，すべてのノードの状態を $s_i = 0$ で，すべてのリンクの状態を $t_{i,j} = 0$ で，それぞれ初期化する．その後，開始ノード o について訪問処理を行う．

訪問処理は visitNode 関数で行う．visitNode 関数では，引数として受け取ったノード i を訪問済みとするため，$s_i = 1$ に更新する．その後，現在地ノードを $c = i$ に更新し，updateAdjNodes 関数を呼び出す．updateAdjNodes 関数では，ノード i に隣接するノード j の状態を $s_j = 2$ に更新することで，ノード v_j を訪問済みノードに隣接するノードとして扱う．

29 行目から 31 行目の onVisit 関数はイベントハンドラであり，ユーザの入力に応じて呼び出される．ここでは，ユーザがノード i を訪問したとき，この関数が呼ばれる．onVisit 関数では，visitNode 関数を呼び出すことで，ノード i に対する訪問処理を行う．visitNode 関数内では，現在地ノード c から訪問先ノード i へのリンク $e_{c,i}$ を通過済みとするため，15 行目の処理で $t_{c,i} = 1$ としている．

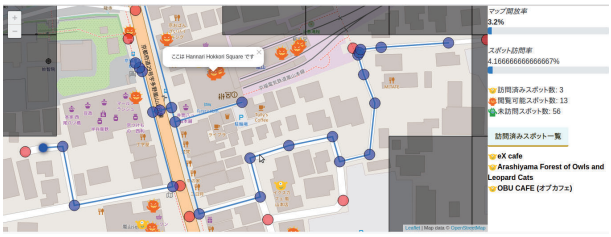


図 4　メッシュの描画の更新．$m_k = 0$ のメッシュは半透明の黒で塗りつぶし，$m_k = 1$ のメッシュは塗りつぶさない．地図画像は Leaflet API を用いて OpenStreetMap の画像を取得した．

4.3　メッシュの隠蔽/解放

リンクの通過/未通過状態に基づきメッシュの隠蔽/解放処理を行う．メッシュk の隠蔽/解放状態 m_k は次式により更新する．

$$m_k = \begin{cases} 1 & \text{if } |\{t_{i,j} \mid t_{i,j} = 1 \wedge t_{i,j} \in M_k\}| \geq 1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

ここで，$t_{i,j} \in M_k$ は矩形領域 M_k にリンク $t_{i,j}$ が含まれることを表す．つまり，矩形領域 M_k に一つでも通過済みのリンクが含まれていれば，$m_k = 1$ となる．

初期状態では，散策エリア全体のメッシュが隠蔽されており，開始ノード o を含むメッシュのみが解放されている状態である．つまり，開始ノード o を含むメッシュの状態のみ，$m_k = 1$ であり，それ以外は $m_k = 0$ である．メッシュの状態の更新はノードを訪問するごとに行う．

Algorithm 1 訪問ノードに基づく更新アルゴリズム．

```
Input: o

c ← NULL
for each v_i ∈ V
    s_i ← 0
end for
for each e_{i,j} ∈ E
    t_{i,j} ← 0
end for
visitNode(o)

function visitNode(i)
    s_i ← 1
    if c ≠ NULL then
        t_{c,i} ← 1
    end if
    c ← i
    updateAdjNodes(i)
end function

function updateAdjNodes(i)
    for each v_j ∈ V
        if s_j = 0 and a_{i,j} = 1 then
            s_j ← 2
        end if
    end for
end function

function onVisit(i)
    visitNode(i)
end function
```

メッシュの隠蔽/解放状態の更新にともない，インタフェース上のマップの描画も更新する．図 4 のように，$m_k = 0$ のメッシュは半透明の黒で塗りつぶし，$m_k = 1$ のメッシュは塗りつぶさない

4.4 スポットの隠蔽/解放

散策エリアには，店舗や神社，景観スポットなどのスポットが配置されている．スポットの状態 u_l は，メッシュの隠蔽/解放状態に応じて次のように更新する．(i)$u_l = 2$（ユーザがスポット l を訪問済みである場合），(ii)$u_l = 1$（ユーザがスポット l を未訪問であり，かつスポット l を含むメッシュk の状態が $m_k = 1$ である場合），(iii)$u_l = 0$（(i)，(ii) 以外の場合）．

$u_l = 0$ のとき，スポット l は非公開であり，マップ上でスポット l の位置は確認できるが，スポット情報を閲覧することはできない．$u_l = 1$ のとき，スポット l は公開であり，マップ上からスポット l をクリックすることで，そのスポットの情報を閲覧することができる．$u_l = 2$ のとき，スポット l は訪問済みであり，同様にスポットの情報を閲覧することができる．

5 おわりに

本研究では，散策時の発見性を促すために，ユーザの訪問経験に応じてマップを隠蔽/解放する動的散策マップシステムを提案した．本システムでは，初期状態はマップのすべての領域を隠蔽しておき，その後，ユーザの訪問経験に応じて，隠蔽されていた領域を解放していく．本稿では，ノードおよびリンクの訪問状態を更新するアルゴリズムを説明し，メッシュおよびスポットの隠蔽/解放処理について述べた．

今後は，被験者実験により本システムの有用性を評価する．具体的には，隠蔽/解放する機能を変えながら比較評価を行うことで，どの要素が散策時の発見性に寄与するか検証する．

謝辞

本研究は JSPS 科研費 19K12567 の助成を受けたものです．ここに記して謝意を表します．

参考文献

[1] Joan Borràs, Antonio Moreno, and Aida Valls. Intelligent tourism recommender systems: A survey. *Expert Systems with Applications*, Vol. 41, pp. 7370–7389, jun 2014.

[2] Chang Shing Lee, Young Chung Chang, and Mei Hui Wang. Ontological recommendation multi-agent for Tainan City travel. *Expert Systems with Applications*, Vol. 36, No. 3 PART 2, pp. 6740–6753, 2009.

[3] Cesar Guzman Laura Sebastia, Inma Garcia, Eva Onaindia. This paper must be cited as : E-Tourism : A tourist recommendation and planning application. *International Journal on Artificial Intelligence Tools*, Vol. 18, No. 5, pp. 717–738, 2009.

[4] Pieter Vansteenwegen, Wouter Souffriau, Greet Vanden Berghe, and Dirk Van Oudheusden. The City Trip Planner: An expert system for tourists. *Expert Systems with Applications*, Vol. 38, No. 6, pp. 6540–6546, 2011.

[5] Arturo Montejo-Ráez, José Manuel Perea-Ortega, Miguel Ángel García-Cumbreras, and Fernando Martínez-Santiago. Otiŭm: A web based planner for tourism and leisure. *Expert Systems with Applications*, Vol. 38, No. 8, pp. 10085–10093, 2011.

[6] Yohei Kurata. CT-Planner2: More Flexible and Interactive Assistance for Day Tour Planning. In *Information and Communication Technologies in Tourism 2011*, pp. 25–37, 2011.

[7] Tomoko Izumi and Yoshio Nakatani. Do turning visited routes in black maps into white promote sightseeing? In *AIP Conference Proceedings*, Vol. 1863, pp. 1–4, 2017.

ARG WI2 No.16, 2020
WI2-2020-05

Unsupervised summarization of arguments toward key point generation with Sentence-BERT-based method

Daiki Shirafuji† Rafal Rzepka†† Kenji Araki††

†Graduate School of Information Science and Technology, Hokkaido University
††Faculty of Information Science and Technology, Hokkaido University
{*d_shirafuji, rzepka, araki*}*@ist.hokudai.ac.jp*

Abstract Many datasets with large number of arguments are used in the field of argument mining. To make a concise summary (key point) from a set of arguments in such datasets, two basic tasks are necessary: match scoring, a step to cluster arguments into similar groups; and key point generation. The previous studies focus on clustering arguments to key points for the first step. However, those studies utilize manually created key points which is time-consuming and costly. To address this problem, we experiment with the classical summarization baselines. We also propose an unsupervised approach for summarizing arguments. In our previous work, we show the Sentence-BERT-large model is most accurate for arguments without labelled data. Therefore, we summarize arguments in our created data with our proposed method. We apply a key phrase extraction method, i.e. EmbedRank, to key point generation using Sentence-BERT. Experimental results show the usefulness of Sentence-BERT for summarizing arguments.

Keywords Natural Language Processing, Argument Mining, Sentence-BERT, Summarization, Debates

1 Introduction

Recently, the area of argument mining has become popular in the field of NLP [1, 2]. This topic spreads from extracting argument components (e.g. claims, premises), predicting persuasiveness of an argument [3, 4], to filling gaps between claims [5, 6], and predicting argument stance (pro/con) [7]. While there are numerous studies on arguments from various perspectives, no method will be practical if users cannot find accurate opinions for their debating topic. Then, how to gather such accurate arguments? There are typically two approaches to collecting arguments; one way is to retrieve them from online debate resources (like *iDebate* or *CreateDebate*), another is to gather arguments from debate spectators who express their opinions [2]. However, both methods have a common problem: it is difficult to read and understand all arguments about a single debate topic because their numbers often exceed hundreds. Reading through such amount of text would be very laborious. In order to solve this problem, Reimers et al. [8] use capabilities of contextualized word embeddings of ELMo [9] and BERT [10] to classify and cluster topic-dependent arguments from Argument Facet Similarity Corpus [11]. They clustered related arguments, however they did not generate summaries of arguments. This problem is addressed in Bar-Haim et al. [12]. In their paper, they defined the key point: usually single sentence describing a set of similar arguments. They generate a large dataset consisting of [argument, key point] pairs. Next, they conduct two experimental steps: *Match Scoring* and *Match Classification*. In *Match Scoring*, they compute a match score for a given [argument, key point] pair, (see Table 1, which shows examples of the data for *Match Scoring*, named ArgKP). In *Match Classification*, their methods discover matching key points for each argument. Although they claim that their research utilizes argument summarization, only the relevance between key points and arguments is calculated, which means their both experiments are almost identical to the above-mentioned research [8]. Furthermore, for *Match Scoring* and *Match Classification*, creating key points is necessary beforehand. This indicates that their methods cannot be used with new debate topics for which key points do not exist. For mapping arguments to key points, they also point out that BERT-large [10] fine-tuned model achieves the best results with supervised learning, and a BERT-large embedding method yields the best results among unsupervised methods. Our previous research [13] showed that Sentence-BERT-large model [14] provides more useful embedding model than other methods in order to embed arguments and key points for *Match Scoring* described above. In that previous work, we assume that one of important factors for summarizing large number of arguments into one key point is to map an argument to the key points, not to

Copyright is held by the author(s).
The article has been published without reviewing.

Topic: **Homeschooling should be banned**

Arguments	Key Points	Stance	Match
education at home could represent a risk since it is not regulated	Homeschools cannot be regulated/standardized	1	1
a parent knows best what is good for their kid	Parents should be permitted to choose the education of their children	-1	0

Table 1: Example of [argument, key point] pairs on a given topic. Stance and Match refer to pro (1) or con (-1), and same meaning (1) or not (0), respectively.

select a key point from one argument. Therefore, we adopt the *Match Scoring*, not the *Match Classification*, utilizing Sentence-BERT.

In this paper, we propose a new task, *Argument Summarization*, for generating a key point from a set of arguments. We summarize arguments utilizing the ArgKP dataset [12], which is explained in Section 3.2, to generate a key point. To the best of authors' knowledge, this is the first attempt of this kind. We utilize the classical, but widely used standard unsupervised models for baselines. Because the size of the dataset is too small for supervised methods, and our previous research shows Sentence-BERT-large model is useful for the ArgKP dataset, we decided to utilize a Sentence-BERT-based unsupervised approach.

The main contributions of this article are:

1. Exploring a new task for *Argument Summarization* toward generating key points for arguments, and providing baselines using various classical unsupervised summarization methods;
2. Proving that Sentence-BERT is useful for summarizing argument when compared with another embedding model;
3. Introducing a new method to extract the exact argument for a key point, i.e. EmbedRank-based method which is based on Sentence-BERT model.

2 Related Works

Argument mining has started with researchers working on argument components for debate topics, discussions or student essays. There are several works on argument persuasiveness [3] or persuasiveness of the whole debate [15] for generating more convincing arguments. In order to conduct research on arguments from several perspectives, various large argument datasets have been recently created [16, 17]. Stab et al. [16] also focus on retrieving arguments from different sources. Their dataset covers about 90% of arguments found in expert-curated lists of arguments from an online debate portal. Similarly, the above-mentioned ArgKP dataset conducted by Bar-Haim et al. [12] also indicates high agreement between expert dataset (key point list) and arguments, achieving Cohen's kappa=0.82. More recently, several researchers have been working on argument facets to identify whether two arguments are similar or not [11, 18]. Others concentrate on examining whether these two arguments have the same facet or not [19]. Their research is similar to mapping arguments to a key point, but they do not calculate similarity between a facet and an argument. Moreover, they do not generate or clarify key points, and in this aspect, their work clearly differs from ours. As for research on mapping arguments to key point, existing research [20, 12] can be given as the most known examples. Boltužić and Šnajder [20] implement argument clustering. They map arguments derived from one online debate portal (*ProCon*) to arguments of other portal (*iDebate*). However, they create only two debate topics with less than 400 arguments. Bar-Haim et al. [12] make a massive dataset including pairs of an argument and a key point labelled with relevancy. They also propose *Match Scoring* and *Match Classification* for the first steps to summarize arguments. *Match Scoring* would be valid for summarization, but *Match Classification* might not be useful for summarizing arguments due to the reasons described in Section 1. Therefore, they do not summarize arguments nor extract any important argument from the argument list. Our work differs from theirs in this aspect.

3 Experimental Setup

In this section, we describe our preparations for *Argument Summarization* task: the task setting, the data for the task, and its evaluation metrics.

3.1 Argument Summarization

In *Argument Summarization*, we summarize a set of similar arguments and generate a key point. We create data

for *Argument Summarization* basically from ArgKP dataset [12], and try to extract the most accurate argument from a set of similar arguments in order to generate a key point. Finally, we evaluate the output with two evaluation metrics.

3.2 Data

We utilize ArgKP dataset for *Argument Summarization*. In order to adapt ArgKP to this task, first we delete the pairs without agreeing arguments and key points, and group arguments into the same key points. As an example for [key point, arguments] pair groups, Table 2 shows a set of arguments which are given for the key point "Parents should be permitted to choose the education of their children" in the debate topic "Homeschooling should be banned". ArgKP also includes some arguments which were paired with more than one key points because they indicate two or more facets. In the dataset, a [key point, arguments] pair group is counted as one data sample. We divide the dataset into 21 train topics (182 groups), and 7 test topics (61 groups).

3.3 Evaluation Metrics

For evaluation, we used two summary evaluation metrics; ROUGE [21] score, and BERTScore [22]. ROUGE is a recall-based metric for fixed-length texts which is based on n-gram co-occurrence. BERTScore computes similarity scores between each token in the original and generated summaries. This metric is more robust to difficult problems such as paraphrases that occur between arguments implying the similar content. Therefore, we use BERTScore in addition to ROUGE. We evaluate our methods with F1 for ROUGE-1, ROUGE-L and BERTScore.

4 Baselines

For *Argument Summarization*, we adopt LexRank and TextRank as classic, but strong extractive baseline methods as comparison with our proposed method.

4.1 LexRank

LexRank was initially proposed by Erkan and Radev [23]. It is a graph-based extractive summarization method for computing relative importance of texts. It constructs a sentence graph whose edge weights (0 / 1) are decided from a cosine similarity. Using LexRank, we extract the highest ranked argument from all arguments. The threshold to decide the edge weights is 0.3, which maximize F1 of BERTScore on the train data.

4.2 TextRank

Page et al. [24] proposed TextRank, a ranking algorithm based on PageRank [25], which is often used in keyword extraction and text summarization. In order to find relevant keywords, TextRank constructs a word/sentence network by looking which words follow one another. A link is created between two words if they follow one another, and link weight increases if these two words occur more frequently adjacent to each other. With TextRank, we extract the highest ranked argument.

5 Proposed Method

This section describes how to utilize EmbedRank to Argument Summarization.

5.1 EmbedRank

This section describes EmbedRank, which we used for summarizing arguments in the previous work. EmbedRank is an unsupervised method with document embeddings for key word extraction [26]. We adapt EmbedRank, which can be used for key words extraction, to order to extract sentences. Moreover, to obtain more accurate sentence embeddings, we propose to use Sentence-BERT [14] instead of Doc2vec [27] and Sent2Vec [28] which were used by [26].

In EmbedRank, there are two parameters; *key size* and λ. *Key size* defines the number of extracted key phrases. When *key size* equals 1, only one key phrase is extracted by EmbedRank. λ is a parameter in the equation for Maximal Marginal Relevance (MMR) [29] which is introduced in EmbedRank. MMR controls the diversity of extracted words and the relevance between documents and extracted words. With a given input query Q, the set S represents documents that are selected as correct answers for Q. S is populated by computing MMR as described in Equation (1),

$$MMR := \arg\max_{D_i \in R \setminus S} \{\lambda * Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j)\} \quad (1)$$

where R is the ranked list of documents retrieved by an algorithm, S represents the subset of documents in R which are already selected, D_i and D_j are retrieved documents, and Sim_1 and Sim_2 are similarity functions. The smaller λ becomes, the diversity of outcomes increases.

5.2 Sentence-BERT

Siamese Neural Networks [30] consist of two neural networks trained with the same weights. These weights are trained on two different input vectors. Reimers and Gurevych proposed Siamese BERT Network and implement-

Topic: **Homeschooling should be banned**

Key Point: **Parents should be permitted to choose the education of their children**

ID	Arguments
1	a parent knows best what is good for their kid
2	home schooling is a good alternative for many. it is economical, safe and and a better environment for some children. parents have a right to choose what is best for their child.
	⋮
27	the parents should be allowed to home school their kids if that's their choice.
28	this is a free country, people have the right to homeschool their children

Table 2: Example of a key point describing a set of arguments on a given topic

ed it as Sentence-BERT[1] (SBERT for short) [14] trained on Natural Language Inference (NLI), SNLI and MultiNLI datasets, in order to create universal sentence embeddings. Following the original Sentence-BERT [14], we use mean-pooling model.

There are other kinds of Sentence-BERT models, i.e. Sentence-BERT-STSb. SBERT-STSb model is also mentioned in the same work [14]. These models are first fine-tuned on the AllNLI dataset, then on the train set of STS benchmark. With this fine-tuning phrase of training process, SBERT-STSb is well suited for measuring semantic textual similarity.

In our proposed model for *Argument Summarization*, we adopt Sentence-BERT-large model because our previous work shows this model is the best performing one.

5.3 Our Proposed Method

First, we set the parameters and input texts as below. The smaller λ becomes, the diversity of outcomes increases, therefore we set λ to 1. Using cosine similarity for the similarity functions, we compute MMR score between an argument *Arg* and a key point *KP*. In our method, MMR score is calculated with the following Equation (2):

$$MMR = \underset{Arg_i \in A \setminus K}{\arg\max}\{cos_{sim}(Arg_i, KP)\}, \quad (2)$$

where A is the set of candidate arguments, and K is the set of extracted key points. A key point KP is the label for our dataset, therefore we average the embeddings of the input arguments, and adopt the averaged embeddings instead of KP. In order to obtain MMR score, Bennani-Smires et al. [26] computed embeddings of documents with Doc2vec and Sent2Vec. In our *Argument Summarization*, we use Sentence-BERT-large, which is the best performing *Match Scoring* embedding method, in addition to these methods. This is because Bennani-Smires et al. [26] used EmbedRank for documents, e.g. the Inspec dataset, and NUS, whereas our target is [key point, arguments] pair groups, which in most cases include only one sentence. As the Sentence-BERT achieves highest scores in many downstream tasks [14], we decided to adopt it. For comparison, we also compute the summarization evaluation scores with Doc2vec-based EmbedRank[2]. The parameter *key size* is set up to 1 in order to obtain the highest ranked argument.

	R1	RL	BERT
LexRank	0.126	0.126	0.872
TextRank	0.126	0.118	**0.880**
EmbedRank w/ Doc2Vec	0.102	0.101	0.868
EmbedRank w/ SBERT-large	**0.132**	**0.131**	0.879

Table 3: F1 score of ROUGE-1, ROUGE-L and BERTScore results of argument summarization. R1, RL, BERT refer to ROUGE-1, ROUGE-L, and BERTScore, respectively.

6 Evaluation Results

Table 3 shows comparison of ROUGE scores and BERT-Score for various *Argument Summarization* methods evaluated on the test data. As shown in Table 3, our method, Sentence-BERT-based EmbedRank, exceeds classical methods considering ROUGE scores, however our approach achieves almost the same BERTScore as TextRank. This means we cannot claim that our method yields a significant differ-

[1] We use Sentence-BERT model available at `https://github.com/UKPLab/sentence-transformers`

[2] We use the pre-trained Doc2vec model trained on the English Wikipedia corpus. This model is available at `https://github.com/jhlau/doc2vec`.

ence from other methods. Regarding the usage of Sentence-BERT, we can say it is efficient for *Argument Summarization* in addition to *Argument Scoring* because the results of Sentence-BERT-based EmbedRank are superior to Doc2vec-based one except for precision of ROUGE-1.

The reason of the small values for ROUGE-1 and ROUGE-L is that there are some cases that several output arguments have the same meaning to the key point (reference) but only few overlapping words between the key point and those arguments.

7 Error Analysis

As can be observed from Table 4, BERTScore in our method shown in Table 4 is lower than others. However, an argument extracted by Doc2vec-based method (see Table 2) seems to contain another semantic opinion, i.e. it is labelled as similar to two key points: "Homeschools can be personalized to the child's pace/needs" and "Homeschooling is often the best option for catering for the needs of exceptional/religious/ill/disabled students." Even BERTScore cannot evaluate such complex texts, so we should consider introducing the human evaluation. This problem may have negative impact not only on evaluation, but also on extraction. To avoid these cases, we consider working on argument facets in the near future.

8 Conclusion

Our work is the first attempt to summarize arguments to one key point. We proposed a new task *Argument Summarization*, whose purpose is to generate a key point from a set of similar arguments.

For the first step of such key point generation, we focused on the existing task, *Match Scoring*, and our previous work [13] which showed that Sentence-BERT-large is the best embedding model for the ArgKP dataset [12]. Then, we utilized Sentence-BERT-large model to *Argument Summarization*. We utilized the ArgKP dataset for *Argument Summarization* task, and extracted a key argument using classical but widely used methods, i.e. LexRank and TextRank. In addition to the standard methods, we also applied Sentence-BERT-based EmbedRank to the *Argument Summarization* task. We also used Doc2vec-based EmbedRank for comparison. As a result, Sentence-BERT-based EmbedRank did not show significant improvement, however Sentence-BERT-based method outperformed Doc2vec-based one. For the future work, we will resolve a problem of arguments with two or more opinions, as they negatively influenced our results.

Finally, we plan to add the scores, which measure argument persuasiveness, to the MMR score described by Equation (2). Whether or not the extracted arguments are sufficiently persuasive is directly related to the impression we get from them. If the argument is not persuasive enough, debate audience will not be convinced. However, if the opposite is true, they will be convinced. For this reason, we are working on scores for measuring an argument persuasiveness according to our previous research on debate outcome prediction [15].

References

[1] Cabrio, E. and Villata, S.: Five years of argument mining: a data-driven analysis. In International Joint Conferences on Artificial Intelligence, Vol. 18, pp. 5427-5433, 2018.

[2] Lawrence, J. and Reed, C.: Argument mining: A survey. Computational Linguistics, 45(4):765–818, 2020.

[3] Habernal, I. and Gurevych, I.: Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1589-1599, 2016.

[4] Persing, I. and Ng, V.: Why can't you convince me? modeling weaknesses in unpersuasive arguments. In International Joint Conferences on Artificial Intelligence, pp. 4082-4088, Melbourne, Australia, 2017.

[5] Boltužić, F. and Šnajder, J.: Fill the gap! analyzing implicit premises between claims from online debates. In Proc. of the Third Workshop on Argument Mining (ArgMining2016), pp. 124-133, 2016.

[6] Habernal, I., Wachsmuth, H., Gurevych, I., and Stein, B.: Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. arXiv preprint arXiv:1802.06613, 2018.

[7] Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., and Slonim, N.: Stance classification of context-dependent claims. In Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 251–261, 2017.

[8] Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., and Gurevych, I.: Classification and Clustering of Arguments with Contextualized Word Embeddings. In Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 567–578, 2019.

[9] Peters, E. M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L.: Deep contextualized word representations. In Proc. of NAACL-HLT, pp. 2227-2237, 2018.

[10] Devlin, J., Chang, W. M., Lee, K., and Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[11] Misra, A., Ecker, B., and Walker, M.: Measuring the similarity of sentential arguments in dialogue. In Proc. of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 276-287, 2016.

[12] Bar-Haim, R., Eden, L., Friedman, R., Kantor, Y., Lahav, D., and Slonim, N.: From arguments to key points: Towards automatic argument summarization. the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4029-4039, 2020.

[13] Shirafuji, D., Rzepka, R., and Araki, A.: Mapping Arguments

Topic: **Homeschooling should be banned**

Methods	Generated Key Points	*BERT*
Original	Homeschools can be personalized to the child's pace/needs	None
LexRank	homeschooling can provide more one on one attention and can allow the child to learn at their own speed.	0.902
TextRank	homeschooling allows parents to teach using methods that are best suited to their child's way of learning	0.903
EmbedRank w/ Doc2Vec	homeschooling provides an opportunity to tailer the curriculum to the particular childs learning style. this may be particularly advantageous when individual learning needs require special consideration.	0.902
EmbedRank w/ SBERT-large	some children are better suited to the homeschool environment and learn better that way.	0.874

Table 4: Generated key point example compared to an original key point. *BERT* refers to F1 in BERTScore between an original key point and a generated key point.

to Key Point: Match Scoring of Arguments using Sentence Embedding and MoverScore without Labelled Data, Web Intelligence and Interaction, November 27-28, 2020.

[14] Reimers, N. and Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019.

[15] Shirafuji, D., Rzepka, R. and Araki, K.: Debate Outcome Prediction using Automatic Persuasiveness Evaluation and Counterargument Relations ", IJCAI Workshop on Linguistic and Cognitive Approaches To Dialog Agents Workshop, 2019.

[16] Stab, C., Daxenberger, J., Stahlhut, C., Miller, T., Schiller, B., Tauchmann, C., Eger, S., and Gurevych, I.: Argumentext: Searching for arguments in heterogeneous sources. In Proc. of the 2018 conference of the North American chapter of the association for computational linguistics:demonstrations, pp. 21-25, 2018.

[17] Ein-Dor, L., Shnarch, E., Dankin, L., et al.: Corpus Wide Argument Mining-A Working Solution. In AAAI, pp. 7683-7691, 2020.

[18] Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., and Gurevych, I.: Classification and clustering of arguments with contextualized word embeddings. In Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 567-578, 2019.

[19] Misra, A., Ecker, B., and Walker, M.: Measuring the similarity of sentential arguments in dialogue. In Proc. of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 276–287, 2016.

[20] Boltužić, F. and Šnajder, J.: Back up your stance: Recognizing arguments in online discussions. In Proc. of the First Workshop on Argumentation Mining, pp.49-58, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[21] Lin, C. Y.: Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pp. 74-81, 2004.

[22] Zhang, T., Kishore, V., Wu, F., and Weinberger, K. Q., and Artzi, Y.: BERTScore: Evaluating Text Generation with BERT. In International Conference on Learning Representations, 2019.

[23] Erkan, G. and Radev, D. R.: Lexrank: Graph-based lexical centrality as salience in text summarization.Journal of artificial intelligence research, 22: pp. 457-479, 2004.

[24] Mihalcea, R. and Tarau, P.: Textrank: Bringing order into text. In Proc. of the 2004 conference on empirical methods in natural language processing, pp. 404-411, 2004.

[25] Page, L., Brin, S., Motwani, R., and Winograd, T.: The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, 1999.

[26] Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., and Jaggi, M.: Simple Unsupervised Keyphrase Extraction using Sentence Embeddings. In Proc. of the 22nd Conference on Computational Natural Language Learning, page 221–229, 2018.

[27] Le, Q. and Mikolov, T.: Distributed representations of sentences and documents. In Proc. of the 31st International Conference on Machine Learning (ICML 2014), pp. 1188–1196, 2014.

[28] Pagliardini, M., Gupta, P., and Jaggi, M.: Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics, 2018.

[29] Carbonell, J. and Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In Proc. of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 335–336, 1998.

[30] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R.: Signature verification using a "siamese" time delay neural network. In Advances in neural information processing systems, pp. 737-744, 1994.

ARG WI2 No.16, 2020
WI2-2020-06

市民意見分析のための複数の属性の定式化と検証

石田 哲也[†,a]　　関 洋平[††,b]

† 筑波大学情報学群知識情報・図書館学類　†† 筑波大学図書館情報メディア系

a) *s1711482@s.tsukuba.ac.jp*　b) *yohei@slis.tsukuba.ac.jp*

概要　行政の政策や接客業のサービスの質を向上させるためには，都市で暮らす市民によるフィードバックが重要となる．本研究では，ソーシャルメディアのつぶやきに現れる多様な市民意見を整理して抽出する手法を提案する．この際，アプレイザル理論を用いた言語学的なアプローチによって意見のタイプをつぶやきに付与することで，対象に着目した市民意見の抽出が可能となる．提案手法では，BERT モデルをファインチューニングすることで意見タイプ，地域依存性，極性といった複数の属性をつぶやきに付与し，これらの属性を利用して市民意見を抽出する．実験では，5 名の実験参加者によって作成されたデータセットを用いており，各属性が BERT モデルのファインチューニングによって高い精度で推定できることを確認した．さらに，推定した属性を利用して時系列ごとの市民意見の出現頻度を分析することで，市民意見と社会情勢や自治体の対応との強い相関が確認できた．

キーワード　意見分析，アプレイザル理論，Twitter，BERT，分散表現

1　はじめに

国や自治体といった行政による政策や，飲食店等の接客業のサービスの質を向上させるためには，実際に都市で暮らす市民の意見を反映させることが重要となる．本研究では，幅広い市民による多数の意見を入手するため，多くのユーザが日頃感じたことを気軽に述べている Twitter からの市民意見の抽出を目的とする．しかし，Twitter における市民意見の種類は多岐にわたり，従来の意見分析研究における肯定や否定といった極性や，喜びや悲しみといった感情の種類のような特定の観点のみでは市民意見の分析には不十分であると考える．

本研究では，Twitter におけるつぶやきを複数の観点から分析することで，市民意見を整理して抽出する手法を提案する．本論文では，横浜市民による「保育園」と「飲食店のテイクアウトサービス」に関するつぶやきを用いて意見分析コーパスを作成し，複数の BERT[1] モデルをファインチューニングすることによって市民意見を分析した結果の有効性について検証する．

2　関連研究

著者の先行研究 [8] では，アプレイザル理論 [7] に基づいた意見分析コーパスを人手で作成，分析した．アプレイザル理論は，言語学の立場から意見分析を行う考え方であり，対象に着目した意見分析を行うことができる．市民意見をどのような形で社会に反映させるのかを検討するにあたって，その意見が何を対象とした意見であるかは重要な観点となる．また，アプレイザル理論では明確な基準によって意見が体系化されているため，客観的な意見分析を行うことが可能となっている．主観的な判断に依存しない，妥当な基準によって意見分析コーパスを作成するにあたり，客観的な分析を行えることは重要と考える．本研究では，つぶやきにアプレイザル理論に基づく意見のタイプを付与したコーパスを作成し，作成したコーパスをモデルの訓練に用いることで Twitter における市民意見のタイプの自動推定を試みる．

栗原ら [5] は，ルールベース手法に基づき，Twitter のつぶやきからの自治体への要望抽出手法を提案した．市民の要望を行政に反映させることは重要だが，市民が日頃抱えている意見は要望のみではなく，より網羅的な意見分析が必要となる．また，ルールベース手法のみを用いて全ての市民意見を抽出することは非常にコストが高くなる．本研究では，深層学習手法に基づき，要望のみにとどまらない多様な市民意見の抽出を試みる．

Devlin et al.[1] によって提案された BERT は，単語間の重み付けに着目することで文脈を考慮した単語や文の分散表現を取得できる．また，自然言語処理分野の深層学習において，大規模なデータによる事前学習と，事前学習済みモデルの各タスクへのファインチューニングを可能にした．さらに，BERT は感情分析タスクにおいても高い性能を誇ることが多くの研究で示されている [2],[9]．本研究では，BERT をファインチューニングすることで複数の分類モデルを構築し，つぶやきを複数の観点から分析する．

3　提案手法

提案手法による Twitter からの自動の意見抽出の流れは，以下の通りである．また，概要を図 1 に示す．

(1) 意見分析コーパスの作成

つぶやきに複数の属性を付与した意見分析コーパスを人手で作成する．本研究では，Twitter における横浜市民のつぶやきを用いて,「保育園」と「飲食店のテイクアウ

Copyright is held by the author(s).
The article has been published without reviewing.

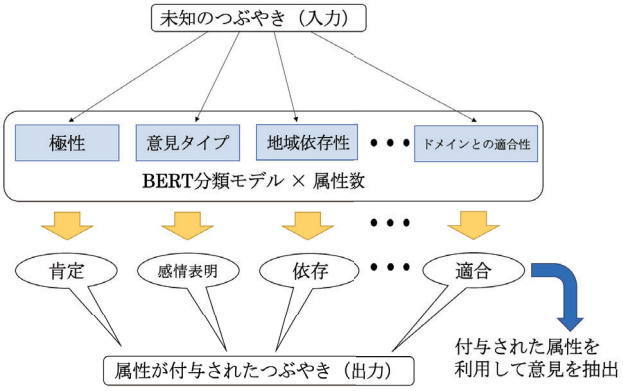


図 1: 提案手法による意見抽出の概要

トサービス」の 2 つのドメインについてコーパスを作成する．

(2) モデルのファインチューニング

作成した意見分析コーパスを用いて，各属性ごとに BERT モデルをファインチューニングする．

(3) 未知のつぶやきからの意見抽出

ファインチューニングした各モデルに未知のつぶやきを入力することで，つぶやきの各属性を予測する．これによって，自動で全てのつぶやきに複数の属性を付与し，これらの属性を組み合わせることで市民意見を抽出する．

本手法では，2 つのドメインを横断して付与する属性と，各ドメインにのみ付与する属性を定義する．各ドメインにのみ付与する属性は，特定の話題との関連性とする．各ドメインにおいて頻出すると考えられる特定の話題については，あらかじめ関連性を判断するモデルを訓練しておき，特定の話題に関連する意見を抽出できるようにする．つぶやきに付与する属性は以下の通りとする．

ドメインを横断して付与

- 極性
  行政や接客業の改善点を探す際には否定的な市民意見が重要となり，飲食店の高評価の口コミのような情報が欲しい際には肯定意見が重要となる．よって，市民意見を活用する場面によって求められる極性は異なると考え，つぶやきの極性を判断する．選択肢は「肯定」,「否定」,「中立」,「意見無し」.
- 意見タイプ
  飲食店が接客を改善する際には店員の振舞に関する意見が重要となり，商品の改良をする際には商品の評価のような意見が重要となるように，意見の対象によって活用の方法は異なる．よって，アプレイザル理論に基づき，対象に着目した意見のタイプを判断する．選択肢は「要求を表す意見」,「自発的感情の表明」,「人間・組織の振舞や行為を対象とした意見」,「事物・事象を対象とした意見」,「意見無し」.
- 中立的な意見タイプ
  著者の先行研究 [8] を参考に，アプレイザル理論には表れないが，Twitter のつぶやきにおいて頻出する意見の種類を判断する．これによって意見の網羅性を高める．選択肢は「推測」,「提案」,「疑問」,「該当無し」.
- 地域依存性
  その意見が都市で暮らす市民特有の意見か，もしくは社会一般的な意見かは，市民意見を社会に反映させる際に重要な要素となる．地域に依存するとは，市や県等の地名や，地域の特定が可能な飲食店等の施設名を含むものと定義する．選択肢は「依存」,「非依存」.
- ドメインとの適合性
  この属性は，単語としては各ドメインに関連するが，意味的には関連のないつぶやきからの意見抽出を避けるために定義する．たとえば，飲食店のテイクアウトサービスについての意見を抽出する際に，「仕事を家に持ち帰りすることになった」といった内容のつぶやきを「今日は美味しいお持ち帰り弁当を買った」といった内容と同様に抽出するのを避けることを目的とする．選択肢は「適合」,「不適合」.
- 投稿主の立場
  同じ内容についての意見であっても，市民の置かれた立場によって意見の方向性は異なる．よって，どのようなユーザによって投稿されたつぶやきであるかを判断する．選択肢は，保育園ドメインが「小さい子を持つ親」,「保育園関係者」,「その他」，飲食店のテイクアウトサービスドメインが「店を利用した人」,「飲食店」,「その他」.

「保育園」ドメインにのみ付与

- 休園・登園自粛との関連性
  新型コロナウイルス感染症によって大きな問題となった，保育園の休園や登園自粛といった話題との関連性を判断する．選択肢は「関連する」,「関連しない」.
- 保育園の定員との関連性
  保育園の定員，具体的には横浜市で大きな問題となっている待機児童問題や，保育園の合否のような話題との関連性を判断する．選択肢は「関連する」,「関連しない」.

「飲食店のテイクアウトサービス」ドメインにのみ付与

- 商品の評価との関連性
  商品の味や量，提供状態等の評価を含むものであるかを判断する．選択肢は「関連する」,「関連しない」.

つぶやき内には複数の意見表現が現れることが多く，意見に直接関連する属性を付与する対象として，つぶや

表 1: 各属性の判定者間一致度（Fleiss の κ 係数）

ドメイン	属性	一致度（A チーム）	一致度（B チーム）
保育園	地域依存性	0.921	0.900
	ドメインとの適合性	0.721	0.707
	休園･登園自粛との関連性	0.847	0.846
	保育園の定員との関連性	0.879	0.865
	投稿主の立場	0.841	0.830
	極性	0.673	0.693
	意見タイプ	0.642	0.635
	中立的な意見タイプ	0.745	0.764
テイクアウト	地域依存性	0.875	0.841
	ドメインとの適合性	0.913	0.835
	商品の評価との関連性	0.828	0.803
	投稿主の立場	0.841	0.830
	極性	0.778	0.766
	意見タイプ	0.722	0.781
	中立的な意見タイプ	0.783	0.754

き全体という単位は大きすぎると考える．そこで本手法では,「極性」,「意見タイプ」,「中立的な意見タイプ」の 3 つの属性については，つぶやきをより細かく分割した，文，あるいは一つだけの意見が含まれる節を単位として属性を付与する．

4　データセット

4.1　データ収集

はじめに，プロフィール情報に基づいて収集した横浜市民の Twitter アカウント 82,583 件のつぶやきを，Twitter の Streaming API を用いて収集した．このつぶやきのうち，2020 年 1 月 1 日から 2020 年 7 月 12 日までの計 29,017,591 件のつぶやきから,「保育園」,「保育士」,「保活」,「待機児童」の単語が含まれるつぶやきを保育園ドメインのつぶやきとして収集し,「持ち帰り」,「テイクアウト」の単語を含むつぶやきを飲食店のテイクアウトサービスドメインのつぶやきとして収集した．リツイートや重複ツイートは取り除いた．

本研究で用いるデータ数は各ドメインに共通で 2,622 件のつぶやきと，それらを文単位に分割したものとした．つぶやきの文単位への分割には，Python のライブラリ，spaCy[1]を用いた．この際，名詞のみで構成される文は意見性を含まないものが多いため，spaCy による分割後に名詞のみで構成された文が生成された場合，つぶやきの先頭以外に現れる文は 1 つ前の文に結合し，先頭に現れる文は 1 つ後ろの文に結合する処理を行った．また，改行，ハッシュタグ，閉じ括弧について，ルールベースの整形処理を行った．

4.2　意見分析コーパスの作成

収集したデータに 3 節で定義した各属性を付与するアノテーション作業を人手で行い，意見分析コーパスを作成した．アノテーション作業は第一著者を含む合計 5 名の実験参加者によって行い，全てのアノテーション結果は多数決によって決定した．なお，全てのデータについて最低 3 名以上の奇数の参加者がアノテーションを担当した．一部のデータで，アノテーション結果が 1-1-1 や 2-2-1 と割れることによって多数決によって結果を決められない場合があったが，このようなデータについては実験参加者間で意見を交換することで，最終的に全ての結果を多数決で決定した．

表 2: 各属性の分類精度（Accuracy）

ドメイン	属性	精度
保育園	地域依存性	0.961
	ドメインとの適合性	0.834
	休園・登園自粛との関連性	0.924
	保育園の定員との関連性	0.959
	投稿主の立場	0.740
	極性	0.717
	意見タイプ	0.692
	中立的な意見タイプ	0.921
テイクアウト	地域依存性	0.885
	ドメインとの適合性	0.962
	商品の評価との関連性	0.905
	投稿主の立場	0.814
	極性	0.828
	意見タイプ	0.815
	中立的な意見タイプ	0.952

アノテーション作業では，はじめに各実験参加者のアノテーション方針を一致させるための訓練を行った．訓練には，保育園ドメイン 250 文（103 件のつぶやき）と飲食店のテイクアウトサービス 250 文（134 件のつぶやき）の計 500 文（237 件のつぶやき）を用い，実験参加者のアノテーション方針が一致したところで訓練を終了した．この際，文単位に分割された文においても複数の意見が含まれると判断した場合は，さらに分割を行うことで，各データにおいて複数の意見表現が現れないように調整した．この分割作業は全員で意見を交換し，過半数の実験参加者間で一致したデータでのみ行った．

訓練終了後に，残りの全てのデータについてのアノテーション作業を行った．第一著者は全てのデータのアノテーションを行い，残りの 4 名を 2 名ずつに分けることで，各 3 名ずつの 2 チームを作り，各チームが残りのデータのうち半数ずつを担当した．この際も過半数の実験参加者間で一致したデータについては，文単位のデータを分割することで同データ内に複数の意見表現が表れないように調整した．Fleiss の κ 係数 [3] を用いた各チームのアノテーションの一致度を表 1 に示す．全ての属性において一致度が 0.6（Substantial Agreement[6]）以上となり，両チームの一致度も近い値であることから，実験参加者によって属性の判定に差異が生まれないことが示された．

[1]https://spacy.io/

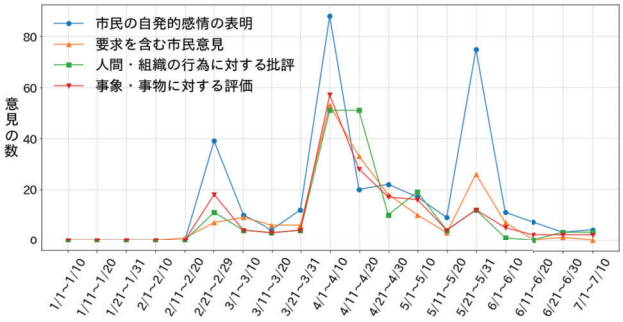


図 2: 保育園休園・登園自粛に関する意見数（正解）

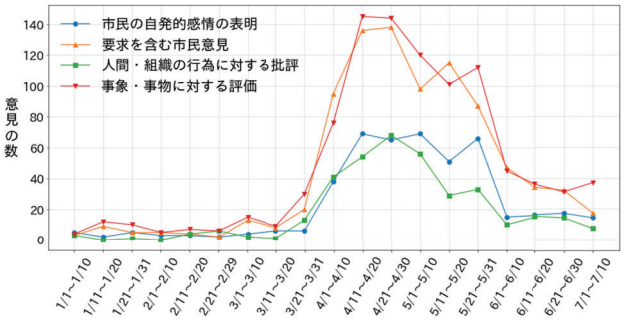


図 4: テイクアウトサービスに関する意見数（正解）

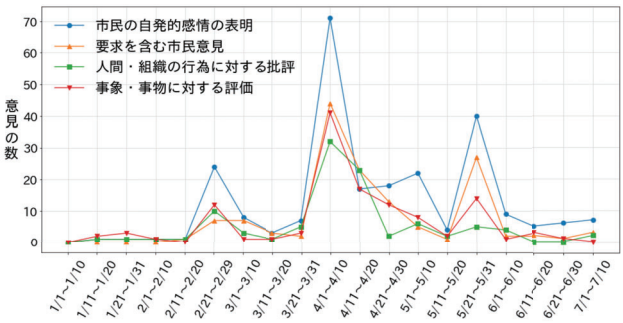


図 3: 保育園休園・登園自粛に関する意見数（推定）

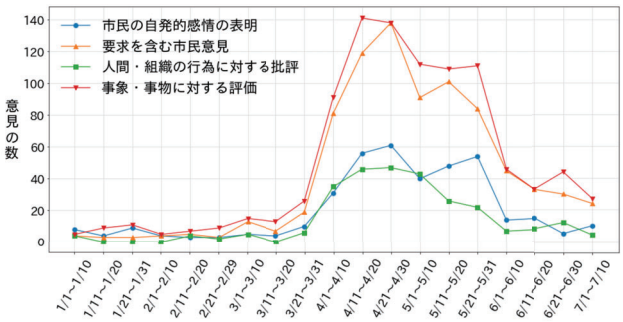


図 5: テイクアウトサービスに関する意見数（推定）

最終的に得られたデータは，保育園ドメインが 2,622 件のつぶやきと 7,916 件の文単位のデータ，飲食店のテイクアウトサービスドメインが 2,622 件のつぶやきと 6,671 件の文単位のデータとなった．

5　実験 1：市民意見の抽出

提案手法による意見抽出の有効性を確かめるため，2 つの実験を行った．本節では市民意見の抽出に関する実験について説明する．

5.1　目的

本実験では，提案手法を用いて未知のつぶやきから市民意見を自動抽出し，各属性の分類精度を評価する．また，BERT のファインチューニングによって高精度で分類可能な属性と，今後改良が必要な属性を明らかにする．

5.2　方法

4 章で作成したコーパスの全てのデータについて，5 分割交差検証を用いて各属性の予測を行い，分類精度（Accuracy）を確認する．

5.3　結果

各属性の分類精度は表 2 の通りである．ほぼ全ての属性について，0.8 以上の分類精度となったが，保育園ドメインの意見タイプ，極性，投稿主の立場については改良の余地がある．

6　実験 2：時系列に着目した市民意見の分析

6.1　目的

ドメインごと，また，意見タイプごとの出現頻度の推移を分析することで，市民意見にどのような推移の傾向があるのかを確認する．また，4 章で作成したアノテーションコーパス（以降正解データと呼ぶ）と自動抽出された市民意見の時系列データの相関を確認することで，推移の傾向を適切に捉えられるかを検証する．

6.2　方法

正解データを用いて，各意見タイプの出現頻度を 10 日間ずつの時系列順に並べ，推移を確認する．また，5.2 節と同様に意見タイプの推定を行い，正解データと同様に出現頻度を 10 日間ずつの時系列順に並べる．今回抽出する意見は，保育園ドメインにおける休園・登園自粛に関連すると推定されたつぶやきの意見タイプと，飲食店のテイクアウトサービスドメインにおける意見タイプを対象とする．また，正解データと推定データの出現頻度のピアソンの相関係数を計算することで，傾向の推定の有効性を確認する．

6.3　結果

保育園休園・登園自粛に関する正解データの時系列順の意見の出現頻度を図 2，推定データの時系列順の意見の出現頻を図 3，飲食店のテイクアウトサービスに関する正解データの時系列順の意見の出現頻度を図 4，推定

表 3: 保育園休園・登園自粛に関する
各意見タイプの出現頻度の正解と推定の相関係数

意見のタイプ	相関係数
自発的感情の表明	0.964
要求を表す意見	0.979
人間・組織の振舞や行為を対象とした意見	0.954
事物・事象を対象とした意見	0.977

表 4: 飲食店のテイクアウトサービスドメインにおける
各意見タイプの出現頻度の正解と推定の相関係数

意見のタイプ	相関係数
自発的感情の表明	0.967
要求を表す意見	0.995
人間・組織の振舞や行為を対象とした意見	0.986
事物・事象を対象とした意見	0.992

データの時系列順の意見の出現頻度を図 5 に示す．また，保育園休園・登園自粛に関する各意見タイプの正解と推定のピアソンの相関係数は表 3，飲食店のテイクアウトサービスドメインにおける各意見タイプの正解と推定のピアソンの相関係数を表 4 に示す．相関係数は全ての意見タイプについて 0.95 を越えたことから，高い相関が確認できた．

7　考察

7.1　実験 1：考察

各属性の分類精度は概ね高い精度となっているが，保育園ドメインの意見タイプは他の属性と比較して精度が低い．原因を考察するため，保育園ドメインにおける意見タイプの各選択肢の適合率，再現率を確認した．結果を表 5 に示す．表 5 から，意見無しデータの再現率が非常に高くなっている．保育園ドメインの意見タイプ属性では，全 7,916 件中 3,953 件，つまり約半数のデータが意見無しとなっている．意見無しデータの再現率が高く，適合率が 0.743 に留まっていることから，多くのデータを多数派である意見無しと予測し，その予測の精度が低いと分かる．よって，少数派ラベルの分類精度向上のための工夫が必要となる．具体的には，多数派ラベルのダウンサンプリングや，損失関数の重み付けが挙げられる．保育園ドメインの極性と投稿主の立場についても，多数派ラベルの再現率が高く適合率が低いことを確認したので，同様の工夫が必要となる．

次に，各 BERT モデルによって推定した属性を付与した未知のつぶやきから，実際に条件を指定することによって抽出された市民意見の例を以下に示す．

まず，保育園ドメインにおいて，地域に依存し，ドメインと適合する，否定的な人物・組織への批評という条件で抽出された意見の例を以下に示す.[2]

[2]本稿に記載されている全てのつぶやきは表現を一部改変している．

表 5: 保育園休園・登園自粛に関する各意見の再現率と適合率

意見のタイプ	再現率	適合率
自発的感情の表明	0.563	0.583
要求を表す意見	0.696	0.755
人間・組織の行為を対象とした意見	0.421	0.607
事物・事象を対象とした意見	0.439	0.589
意見無し	0.864	0.743

- 同じ横浜市でも保育園によって全然対応違うんだね。子連れ出勤で子供の保育料六万が給料から引かれるの恐ろしすぎません?
- 中にはあずけざるを得ない保護者もいて時にはクレームの対応もするんだろうな... 国なり県なり市なりが先頭に立ってくれよ。全責任を保育園に丸投げするなよ。それどころか、命を守ろうとしてる行動を阻むって何。#横浜市

次に，飲食店のテイクアウトサービスドメインにおいて，店を利用した人による，地域に依存し，ドメインと適合する，商品の評価に関する肯定的な事物・事象の評価という条件で抽出された市民意見の例を以下に示す.

- 今日は #ワイン中華カントナ のテイクアウト 🐼 可愛い見た目に反して 中身は本格中華というギャップがたまらない ^^ 大きなカニ爪もまた旨い!
- 磯子区の麺屋づかちゃんで汁なしとお土産チャーシューをテイクアウト　ホロホロの豚に麺と絡む汁、玉ねぎのシャキシャキ感がいい さいこ〜です 😆

実際に抽出された意見を確認したところ，条件通りの意見をうまく抽出できることが分かった．保育園ドメインにおいては，地域に依存すると指定した上で否定的な人物・組織への批評意見という条件を組み合わせることで，横浜市民ならではの保育園に関する批評意見を抽出できている．飲食店のテイクアウトサービスドメインにおいては，地域依存性有りと指定することで，どの飲食店についての意見であるかを推定可能であることを確認した．また，飲食店を利用した人による肯定的な評価という条件を組み合わせているため，飲食店のレビューサイトにおける高評価レビューのような意見を未知のつぶやきから抽出可能であることを明らかにした．

7.2　実験 2：考察

抽出した市民意見の投稿時期について分析し，市民意見の出現頻度と社会情勢や社会情勢への自治体の対応との相関を確認した．

保育園休園・登園自粛に関する意見は，新型コロナウイルス感染症が国内で流行し始めた 2 月下旬に現れ始め

ている．横浜市が緊急事態宣言発令後の保育園の休園に関する方針を発表した 4 月の上旬と，緊急事態宣言解除後の保育園の運営に関する方針を発表した 5 月下旬にも，横浜市民の保育園休園・登園自粛に関する意見が急激に増加している．飲食店のテイクアウトサービスに関する意見では，新型コロナウイルス感染症の流行が国内で本格化し，多くの飲食店がテイクアウトサービスを開始し始めた 4 月上旬に急激に意見が増加している．

また，保育園の休園・登園自粛に関する意見については，自治体が政策を発表した際に，市民は平常時よりも多くの要求意見を述べるという傾向があることが分かった．そこで，横浜市が緊急事態宣言発令後も保育園の運営を続けるという方針を発表した 4 月上旬のつぶやきから，保育園の休園・登園自粛に関連し，要求を含む意見という条件で市民意見を抽出した．結果として，以下のような意見を抽出することができた．

- いっそのこと休園してくれ。横浜市は原則開園って。。。保育士のみなさんもつらいだろうに
- 横浜の保育園は原則開園らしいけど、家庭で保育できるなら保育料減額してくれるみたい。 休みたいよ…

以上から，特定の時期における意見を明示的に抽出することで，保育園休園・登園自粛に関する自治体の発表した方針といった，社会情勢への自治体の対応と関わりのある意見を抽出できることが明らかになった．

時系列に着目した相関係数を見ると，保育園ドメインにおける人物・組織の行為や振る舞いの批評や，テイクアウトドメインにおける自発的感情の表明は値が低くなっている．これらの意見タイプは，他の意見タイプと比較して再現率が低い．現在，モデルの性能は分類精度のみによって評価しているが，今後は各選択肢の再現率についても考慮した訓練を検討していきたい．

8　おわりに

本研究では，Twitter のつぶやきを複数の観点から分析することで，条件を指定して市民意見を自動抽出する手法を提案した．つぶやきに付与する属性の分類精度については，多くの属性が BERT モデルのファインチューニングによって高精度で分類可能であることが分かったが，アプレイザル理論に基づく意見タイプについては今後精度向上のための工夫を検討する必要があることが分かった．また，各属性の情報を用いて，実際に未知のつぶやきから市民意見を抽出したところ，地域依存性や極性，意見タイプ等を捉えた意見が意図した通りに自動で抽出できることを明らかにした．

時系列に着目した意見分析においては，時系列ごとの市民意見の出現頻度から社会情勢や自治体の対応との関わりを読み取ることが可能であり，また，時期を指定して市民意見を抽出することによって，特定の社会情勢への自治体の対応と関わりがある市民意見を抽出することが可能である．

今後の課題として，関係する 2 つの属性を同時に学習することで双方の分類精度を向上させる手法である Multi-task learning[4] を用いて各属性の分類精度の向上を試みることや，意見抽出の対象となる都市の拡張を行い，特定の地域に依存しない市民意見の抽出を実現することが挙げられる．

謝辞

本研究の一部は，科学研究費補助金基盤研究 B（課題番号 19H04420）の助成を受けて遂行された．

また，横浜市政策局共創推進課の関口昌幸さんには，行政視点からの貴重な助言を頂きました．ここに深く感謝します．

参考文献

[1] Devlin, J. Chang, M. Lee, K and Toutanova, K.: BERT: pretraining of deep bidirectional transformers for language understanding, Proc. of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171-4186, 2019.

[2] Du, C. Sun, H. Wang, J. Qi, Q and Liao, J.: Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis, Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4019-4028, 2020.

[3] Fleiss, J. L.: Measuring nominal scale agreement among many raters, Psychological Bulletin, Vol. 76, No. 5, pp. 378-382, 1971.

[4] Hashimoto, K. Xiong, C. Tsuruoka, Y and Socher, R.: A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks, proc. of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1923–1933, 2017.

[5] 栗原理聡，佐々木彬，松田耕史：Twitter を利用した地域毎の要望抽出，人工知能 学会全国大会論文集，Vol. 29, pp. 1-4, 2015.

[6] Landis, J. R. and Koch, G. G.: The Measurement of Observer Agreement for Categorical Data, Biometrics, Vol. 33, No. 1, pp. 159-174, 1977.

[7] Martin, J. R. and White, P. R. R.: The Language of Evaluation: Appraisal in English, Palgrave Macmillan, 278p, 2005.

[8] 関洋平：コミュニティQA における意見分析のためのアノテーションに関する一 検討，自然言語処理，Vol. 21, No. 2, pp. 271-299, 2014.

[9] Yin, D. Meng, T and Chang, K.: SentiBERT: A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics, Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3695-3706, 2020.

ARG WI2 No.16, 2020
WI2-2020-07

Mapping arguments to key point: Match Scoring of arguments using sentence embedding and MoverScore without labelled data

Daiki Shirafuji† Rafal Rzepka†† Kenji Araki††

†Graduate School of Information Science and Technology, Hokkaido University
††Faculty of Information Science and Technology, Hokkaido University
{*d_shirafuji, rzepka, araki*}*@ist.hokudai.ac.jp*

Abstract There might be hundreds of arguments for a single debate topic. and it is usually difficult to read them all. Therefore, some of the existing studies have attempted to summarize and aggregate these arguments manually into about 14 types of content (key point) in order to make them easier to understand. Some researchers pair key points with arguments and use BERT to identify whether an argument and a key point are matching or not. However, the previous study averaged word vectors to compute sentence vectors, which resulted in the unbalance of important word information. This means that the important words should be weighted more than other words, but the previous study does not consider about that. To solve this problem, we implemented a sentence embedding model called Sentence-BERT, which is fine-tuned on NLI dataset with BERT. The results are better than the existing SOTA method, i.e. BERT. Furthermore, to compare each words between two sentences, we introduce MoverScore, which yield the best results.

Keywords Natural Language Processing, Argument Mining, Sentence-BERT, MoverScore, Debates

1 Introduction

Arguments in debates or discussions are broadly studied in NLP area [1, 2]. Research field related to arguments is called argument mining, and this field has been increasingly famous in this ten years. In argument mining, we basically extract the argument structures from unstructured texts or utterances [3]. These attempts for detecting argument components (e.g. claims, premises) have been clarifying where an author (or a speaker) mainly claims, where is an evidence for their claim, and where they show an example for their argument. Nowadays, researchers are getting better at distinguishing above-mentioned components, therefore targets of the area of argument mining are gradually moved on to different topics. Argument mining researchers have focused on predicting persuasiveness of an argument [4, 5], filling gaps between claims [6, 7], and predicting argument stance (pro/con) [8, 9]. These works have benefited from enormous amount of argument data from debates, discussions, speeches, or essays.

However, studies on aggregating such numerous arguments seem to be neglected. Numerous arguments make it hard to know which argument is better to read. Although there are various and numerous studies in the field of argument mining, no method will be practical if users cannot find accurate opinions for their debating topic. We have to narrow down the target argument, i.e. summarize a lot of arguments into one core argument. To address this problem, we also need plentiful arguments, and a new question arises; how to gather arguments? There are typically two approaches to aggregating arguments; one way is to retrieve them from online debate resources (like *iDebate* or *CreateDebate*), another is to gather arguments from debate spectators who express their opinions [2]. Using both methods, the number of gathered arguments in a single debate topic often exceed hundreds or more. To avoid laborious trouble, reading through such amounts of text, one need to work on summarizing arguments.

In order to address this problem, Reimers et al. [10] use capabilities of contextualized word embeddings of ELMo [11] and BERT [12] to classify and cluster topic-dependent arguments from Argument Facet Similarity Corpus [13]. They clustered related arguments, however they did not generate summaries of arguments. This problem is mentioned in Bar-Haim et al. [14], who defined the key point: usually single sentence describing a set of similar arguments, i.e. a summary of arguments. A large dataset consisting of [argument, key point] pairs is generated. Next, two experimental steps are conducted: *Match Scoring* and *Match Classification*. In *Match Scoring*, a match score for a given [argument, key point] pair is computed, (see Table 1, which

Copyright is held by the author(s).
The article has been published without reviewing.

Topic: **Homeschooling should be banned**

Arguments	Key Points	Stance	Match
education at home could represent a risk since it is not regulated	Homeschools cannot be regulated/standardized	1	1
a parent knows best what is good for their kid	Parents should be permitted to choose the education of their children	-1	0

Table 1: Example of [argument, key point] pairs for a given topic. Stance and Match refer to pro (1) or con (-1), and the same meaning (1) or not (0), respectively.

shows examples of the dataset for *Match Scoring*, named ArgKP). In *Match Classification*, their methods discover matching key points for each argument. For these tasks, they describe that BERT-large [12] fine-tuned model is achieving the best results in supervised learning, and a BERT-large embedding method is the best among unsupervised methods. Although Bar-Haim et al. claim that their research utilizes argument summarization, only the relevance between key points and arguments is calculated, which means their both experiments are almost identical to the above-mentioned research of Reimers et al. [10]. Furthermore, for *Match Scoring* and *Match Classification*, creating key points is necessary beforehand. This indicates that their methods cannot be used with new debate topics for which key points do not exist.

The mapping step is useful when generating key point from a set of arguments. *Match Scoring*, which is explained above, judges whether a given argument and a given key point are related or not. Following this *Match Scoring* step, it is possible to create a set of arguments for generating a key point.

In this paper, we address the *Match Scoring* task for key point generation with the ArgKP dataset [14], which is explained in Section 3.2. We assume that the important factor for summarizing large number of arguments into one key point is to map an argument to key points, not to select a key point from one argument. Therefore, we adopt the *Match Scoring*, not the *Match Classification*, for our target task in this paper. The previous research [14] experimented with both supervised and unsupervised methods, however we tackle the task with unsupervised approach because the size of the data is too small when we utilize the ArgKP dataset for key point generation. The previous unsupervised approaches only averaged word vectors for computing a sentence vector, so that the weight of important word information is identical with other words during sentence vector calculation. To address this problem, we try to solve it applying two methods: Sentence-BERT [15] and MoverScore [16]. We hypothesize that sentence embedding models could be a good solution to represent a whole sentence in a vector, and that output vector will not be able to represent significant words in an argument. Therefore, we utilize variations of Sentence-BERT model, which is fine-tuned with NLI dataset and is able to represent sentence embeddings. We also utilize MoverScore, a metric to measure text similarity using Word Mover's Distance [17] in order to calculate similarity between an argument and a key point on the word level. To the best of authors' knowledge, this is the first attempt to address the important word problem with sentence embedding models and MoverScore metric for the *Match Scoring* task.

The main contributions of this article are:

1. Pointing out that existing methods which average word embeddings to compute sentence embeddings, cannot weight important word words in an argument;
2. Applying two methods to put more weight on significant words and less on non-significant words, i.e. Sentence-BERT and MoverScore;
3. Experimentally proving that both Sentence-BERT and MoverScore methods outperform the existing unsupervised methods, and the MoverScore approach yield the best result.

2 Related Works

This section describes works related to this study. First section explains the history of the field, and similar topics in the area of argument mining. The second section presents approaches to measure text similarity.

2.1 Argument Mining

Argument mining has started with researchers working on argument components [3] for debates, discussions or student essays. After argument structure had been gradually

solved, several researchers began to work on argument persuasiveness [4] or persuasiveness of the whole debate [18] in order to generate more convincing arguments. Stance prediction is also active research topic on argument mining. Several works estimate whether a given argument is following a given debate topic or not (pro/con) [8, 9]. In order to conduct research on arguments from several perspectives, various large argument datasets have been recently created [19, 20]. Stab et al. [19] also focus on retrieving arguments from different sources. Their dataset covers about 90% of arguments found in expert-curated lists of arguments from an online debate portal. Similarly, ArgKP dataset conducted by Bar-Haim et al. [14] also indicates high agreement between expert dataset (key point list) and arguments, achieving Cohen's kappa=0.82.

More recently, Misra et al. [13] attempt to measure a similarity of arguments in debate portal. They provide Argument Facet Similarity Corpus. The purpose of their work is to identify whether an argument has the same facets as other arguments in across multiple conversations. Several researchers are following their topic with logical rules or contextual embedding models [21, 22]. These works are similar to mapping arguments to key points, but they do not calculate similarity between a facet and an argument. Moreover, they do not clarify key points, and in this aspect, their work clearly differs from ours.

As for mapping arguments to key points, i.e. *Match Scoring* and *Match Classification* tasks, existing works [23, 14] can be given as most known examples. Boltužić and Šnajder [23] implement argument clustering. They map arguments derived from one online debate portal (*ProCon*) to arguments of other portal (*iDebate*). However, they create only two debate topics with less than 400 arguments. Bar-Haim et al. [14] make a massive dataset including pairs of an argument and a key point labelled with relevancy. They also propose *Match Scoring* and *Match Classification* as the first steps to summarize arguments.

2.2 Text Similarity

NLP researchers recently utilize vectors of words / documents for measuring similarity between texts. Today, we usually compute word vectors with word embedding, sentence embedding, or language models. One of the most well-known algorithms, word2vec [24] is a pre-trained skip-gram or Continuous Bag-of-Words (CBOW) model to represent words as vectors. Following the pre-trained word embedding models, sentence / document embedding models have been also developed. For example, Skip-Thought Vector [25] is based on a sentence encoder that predicts the surrounding sentences of a given sentence. Skip-Thought is based on an encoder-decoder model, its encoder maps words to a sentence vector and its decoder generates the surrounding sentences. Sentence Transformers [15] are sentence embedding models for English texts based on siamese / triplet networks [26, 27]. They insert language models (like BERT [12]) into the network of siamese / triplet networks. They call BERT-based siamese / triplet networks as "Sentence-BERT."

There are also some other methods to compute text similarity. Word Mover's Distance [17] algorithm calculates the distance between texts with word2vec, which introduces Earth Mover's Distance [28] into NLP field. Inspired by Word Mover's Distance, Zhao et al. [16] investigate encoding systems to devise a metric that shows a high correlation with human judgment of text quality in summarization or machine translation tasks. They name their metric MoverScore, whose details are described in Section 5.3.

Our proposed methods are based on Sentence-BERT and MoverScore, which are described above.

3 Experimental Setup

In this section, we describe the preparation for *Match Scoring* task; the task setting, and the data for our experiments.

3.1 Match Scoring

In *Match Scoring*, we map arguments to a key point which may have similar meaning to them, and compute a match score for [argument, key point] pairs. As examples of matching and not matching pairs, in Table 1, we show pairs, their stance to a given topic, and the manually annotated matching label. We try to identify exactly whether such arguments and key points are similar or not. We evaluate match scores with accuracy, precision, recall, and F1.

3.2 Data

We utilize ArgKP dataset [14] for *Match Scoring*. This dataset consists of about 24,000 [argument, key point] pairs which are labeled as follows; (1) whether an argument and a key point describe the same content or not; and (2) whether an argument represents pro side or con side. To the best of authors' knowledge, ArgKP is the largest dataset usable for the *Match Scoring*. It is based on IBM-Rank-30k dataset [29] which contains 30,497 arguments annotated for their quality. Using the quality-indicating labels, Bar-Haim et al. [14] filter out unclear arguments whose quality score is

lower than 0.5 or polarity score is lower than 0.6. After the arguments are filtered, they generate key points, and pair an argument with a key point. They annotate a [argument, key point] pair with the label whether they are similar to each other or not. For each stance in one topic, 6.75 key points were generated on average, and there are 378 key points in total. With 24,000 [argument, key point] pairs in ArgKP dataset, we divide the dataset into 7 test topics and 21 train topics, following experimental setup of the previous research [14].

4 Baseline Methods

For *Match Scoring* task, we adopt BERT-based and Word-2Vec with WMD methods for comparing with our proposed approaches described in the Section 5.

4.1 BERT-large Fine-tuned

Bar-Haim et al. experimentally showed that fine-tuned BERT-large model yielded the best score in the several supervised methods [14]. Note that their approach is supervised, unlike the others.

4.2 BERT-large Embedding

For comparison with other unsupervised approaches, we choose BERT-large [12] as the SOTA system for *Match Scoring* [14]. When we embed an argument and a key point using BERT-large, we examine averaged word embeddings as their embeddings. If the cosine similarity between an argument and a key point is above a threshold decided with the train data, the methods regard the pair as relevant.

4.3 Word2Vec without WMD

When we embed an argument and a key point using Word2Vec, we examine averaged word embeddings as sentence embeddings. If the cosine similarity between an argument and a key point is above a threshold decided with the train data, the methods regard the pair as relevant.

4.4 Word2Vec with WMD

We experiment with Word Mover's Distance (WMD) [17] to compare with MoverScore-based method. We compute distance between an argument and a key point with WMD. If the dissimilarity between an argument and a key point is BELOW a threshold decided with the train data, the methods regard the pair as relevant.

5 Proposed Methods

This section describes our proposed methods for *Match Scoring* task, i.e. Sentence-BERT, Sentence-BERT fine-tuned on STS benchmarks, and MoverScore approaches. All thresholds for similarity function are acquired from the train data for each method. The detailed way to decide thresholds is following;

1. Calculating F1 score for the positive (matching) class for each threshold by clustering the train data;
2. Obtaining the thresholds which maximize the F1 score for the train data.

As a result, the thresholds of Sentence-BERT-base, Sentence-BERT-large, Sentence-BERT-STSb-base, Sentence-BERT-STSb-large, and MoverScore are 0.64, 0.71, 0.58, 0.56, and 0.12 respectively.

5.1 SBERT

Siamese Neural Networks [30] consist of two neural networks trained with the same weights. These weights are trained on two different input vectors. Reimers and Gurevych [15] proposed Siamese BERT Network and implemented it as Sentence-BERT[1] (SBERT for short) trained on Natural Language Inference (NLI), SNLI and MultiNLI datasets, in order to create universal sentence embeddings. Following the original Sentence-BERT, we use the mean-pooling model.

5.2 SBERT-STSb

SBERT-STSb model is also proposed in [15]. These models are first fine-tuned on the AllNLI dataset, then on the train set of STS benchmark. With this fine-tuning phrase of training process, SBERT-STSb is well suited for measuring semantic textual similarity.

5.3 MoverScore

MoverScore [16] is a robust evaluation metric to evaluate summarization or machine translation tasks. We adopt this scoring metric to calculate similarity between an argument and a key point for calculating distances between words in each sentence. MoverScore is based on WMD or Sentence Mover's Distance (SMD) [31]. To measure semantic distance, n-gram (n=1, 2) is utilized in addition to word distances. For embedding model calculating word distances, Word2Vec, ELMo, and BERT-base are adopted. BERT-base is fine-tuned on one among following datasets: MNLI, QANLI, or QQP. Furthermore, the word representations from ELMo and BERT-base are aggregated with Power Means or Routing Mechanism. ELMo and BERT output different vectors from each layer, therefore they aggregate vectors (all three layers for ELMo, and the final five layers for BERT-base).

[1] We use Sentence-BERT model available at https://github.com/UKPLab/sentence-transformers

	Acc	P	R	F1
BERT-large Fine-tuned [14]	0.868	0.685	0.688	0.684
BERT-large Embedding [14]	0.660	0.319	0.550	0.403
Word2Vec w/o WMD	0.670	0.290	0.501	0.368
Word2Vec w/ WMD	0.493	0.218	**0.635**	0.324
SBERT-base	0.682	0.317	0.571	0.408
SBERT-large	0.728	0.354	0.513	0.419
SBERT-STSb-base	0.713	0.341	0.538	0.418
SBERT-STSb-large	0.706	0.335	0.544	0.415
MoverScore	**0.747**	**0.388**	0.554	**0.457**

Table 2: Results of each clustering embedding method. Acc, P, and R refer to accuracy, precision, and recall, respectively.

We use the best metric for summarization tasks given in [16], i.e. uni-gram, BERT-base, MNLI, Power Means, and WMD, for our proposed method.

6 Evaluation Results

The results of accuracy, precision, recall, and F1 score on the test dataset are shown in Table 2. These are results of experiments with the test data using thresholds learned from the train data. As shown in Table 2, SBERT-large achieves higher scores in accuracy, precision, and recall than existing unsupervised SOTA, i.e. BERT Embedding. While SBERT-large achieves a relatively good score, versions of SBERT-STSb do not exceed SBERT-large even though they have been reported as having superior capability for measuring semantic similarity [15]. For this reason, before the experiment, we assumed that SBERT-STSb would exceed SBERT score. STS benchmark and other NLI datasets contain daily conversation-like data. Our intuition is that such data might be unsuitable for debate topics. In future, debate or argument-oriented BERT model could be used instead of the standard BERT [20].

In addition to SBERT, MoverScore yields the best scores except recall. MoverScore relies on soft-alignment (many-to-one), so it allows to map semantically related words in one text to the respective word in another text. This is, in our opinion, why the MoverScore performed the best. Regarding recall score, the problem seems to lay in the dataset bias. As we have already mentioned in Section 3.2, more non-matching data exist than matching ones. This is most probably the reason why the recall value in Word2Vec with WMD is higher than that in MoverScore.

7 Error Analysis

Within the topic "We should ban the use of child actors," argument "the use of child actors exploits a child and can have negative effects on them" and key point "Being a performer harms the child's education" represent different content. However, the similarity between them is high (0.9140). The reason for this may be that both of sentences describe negative effects on child actors, so the method incorrectly classified them as representing the same content. Both argue the negative effects, but from different facets. It can be assumed that our model will obtain better results if it can identify the difference from a perspective of detecting a different facet of a discussed problem. This problem occurs in both of our methods, SBERT and MoverScore. We plan to treat it in near future.

8 Conclusion

Our work is focused on the existing task, *Match Scoring* for the first step toward key point generation. We point out that the existing methods, averaging word embeddings to compute sentence embeddings, cannot weight important words in an argument, therefore we propose to utilize Sentence-BERT model and MoverScore metric for the task, and both of our approaches outperformed the state of the art system (BERT-large) on the ArgKP dataset [14]. This could be due to the fact that our methods weight important words which is not done in other studies. From these results, we can say the sentence embedding models and NLI dataset, which is used for fine-tuning / training Sentence-BERT and MoverScore metric, are useful for the *Match Scoring* task, so we plan to utilize other sentence transformers, like Sentence-RoBERTa which is also fine-tuned on NLI dataset, to improve our results. MoverScore is based on Word Mover's Distance, which means word significance treats A and word meaning as the same vactor. To solve this problem, we are going to introduce Word Rotator's Distance, which can represent the importance and the meaning separately.

For the next step, we will attempt to generate key points from argument set as mentioned in Section 1. We plan to use Maximal Marginal Relevance (MMR), which can identify important words in texts, for summarizing arguments in order to generate key points.

References

[1] Cabrio, E. and Villata, S.: Five years of argument mining: a data-driven analysis. In International Joint Conferences on Artificial Intelligence, Vol. 18, pp. 5427-5433, 2018.

[2] Lawrence, J. and Reed, C.: Argument mining: A survey. Computational Linguistics, 45(4):765–818, 2020.

[3] Aharoni, E., Polnarov, A., Lavee, T., Hershcovich, D., Levy, R., Rinott, R., Gutfreund, D., and Slonim, N.: A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics. In Proc. of the First Workshop on Argumentation Mining, Association for Computational Linguistics, pp. 64–68, 2014.

[4] Habernal, I. and Gurevych, I.: Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers), pp. 1589-1599, 2016.

[5] Persing, I. and Ng, V.: Why can't you convince me? modeling weaknesses in unpersuasive arguments. In International Joint Conferences on Artificial Intelligence, pp. 4082-4088, Melbourne, Australia, 2017.

[6] Boltužić, F. and Šnajder, J.: Fill the gap! analyzing implicit premises between claims from online debates. In Proc. of the Third Workshop on Argument Mining (ArgMining2016), pp. 124-133, 2016.

[7] Habernal, I., Wachsmuth, H., Gurevych, I., and Stein, B.: Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. arXiv preprint arXiv:1802.06613, 2018.

[8] Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., and Slonim, N.: Stance classification of context-dependent claims. In Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Vol. 1, Long Papers, pp. 251–261, 2017.

[9] Popat, K., Mukherjee, S., Yates, A., Weikum, G.: STANCY: Stance Classification Based on Consistency Cues. In Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 6414-6419, 2019.

[10] Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., and Gurevych, I.: Classification and clustering of arguments with contextualized word embeddings. In Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 567–578, 2019.

[11] Peters, E. M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L.: Deep contextualized word representations. In Proc. of NAACL-HLT, pp. 2227-2237, 2018.

[12] Devlin, J., Chang, W. M., Lee, K., and Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[13] Misra, A., Ecker, B., and Walker, M.: Measuring the Similarity of Sentential Arguments in Dialogue. In Proc. of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 276-287, 2016.

[14] Bar-Haim, R., Eden, L., Friedman, R., Kantor, Y., Lahav, D., and Slonim, N.: From arguments to key points: Towards automatic argument summarization. the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4029-4039, 2020.

[15] Reimers, N. and Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019.

[16] Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., and Eger, S.: MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 563-578, 2019.

[17] Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In International conference on machine learning, pp. 957-966, 2015.

[18] Shirafuji, D., Rzepka, R. and Araki, K.: Debate Outcome Prediction using Automatic Persuasiveness Evaluation and Counterargument Relations ", IJCAI Workshop on Linguistic and Cognitive Approaches To Dialog Agents Workshop, 2019.

[19] Stab, C., Daxenberger, J., Stahlhut, C., Miller, T., Schiller, B., Tauchmann, C., Eger, S., and Gurevych, I.: Argumentext: Searching for arguments in heterogeneous sources. In Proc. of the 2018 conference of the North American chapter of the association for computational linguistics:demonstrations, pp. 21-25, 2018.

[20] Ein-Dor, L., Shnarch, E., Dankin, L., et al.: Corpus Wide Argument Mining-A Working Solution. In AAAI, pp. 7683-7691, 2020.

[21] Amgoud, L., and David, V.: Measuring similarity between logical arguments. In Sixteenth International Conference on Principles of Knowledge Representation and Reasoning, 2018.

[22] Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., and Gurevych, I.: Classification and Clustering of Arguments with Contextualized Word Embeddings. In Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 567-578, 2019.

[23] Boltužić, F. and Šnajder, J.: Back up your stance: Recognizing arguments in online discussions. In Proc. of the First Workshop on Argumentation Mining, pp.49-58, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[24] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J.: Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pp. 3111-3119, 2013.

[25] Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S.: Skip-thought vectors. In Advances in neural information processing systems, pp. 3294-3302, 2015.

[26] Hadsell, R., Chopra, S., and LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2, pp. 1735-1742, 2006.

[27] Hoffer, E., and Ailon, N.: Deep metric learning using triplet network. In International Workshop on Similarity-Based Pattern Recognition, pp. 84-92, 2015.

[28] Rubner, Y., Tomasi, C., and Guibas, L. J.: The earth mover's distance as a metric for image retrieval. In International journal of computer vision, Vol. 40(2),pp. 99-121, 2000.

[29] Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., and Slonim, N.: A large-scale dataset for argument quality ranking: Construction and analysis. In AAAI-20, 2020.

[30] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R.: Signature verification using a "siamese" time delay neural network. In Advances in neural information processing systems, pp. 737-744, 1994.

[31] Clark, E., Celikyilmaz, A., Smith, N. A.: Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2748-2760, 2019.

ARG WI2 No.16, 2020
WI2-2020-T1

招待パネル:
情報系のオンライン演習・講義における工夫:
Like duck's webbed feet in underwater

伊藤貴之 (お茶の水女子大学), 中村聡史 (明治大学)

松田昌史 (NTT コミュニケーション科学基礎研究所).

村上綾菜 (お茶の水女子大学), 杉原太郎(東京工業大学)

概要　本パネルディスカッションは, 2020 年度のコロナ禍で取り組まれた情報系のオンライン講義, ないしはオンライン演習における工夫を, 実施した教員の立場, 学生の立場から議論する. また, 紹介された講義・演習における営為を社会心理学の知見に基づき紐解いていく.

ARG WI2 No.16, 2020
WI2-2020-08

LexRankを用いた小説文章からの自動要約手法の検討

安武 凌[†,a]　野中 健一[††,b]　岩井 将行[†,c]

† 東京電機大学大学院未来科学研究科　†† 立教大学文学部

a) *ryo@cps.im.dendai.ac.jp*　c) *iwai@cps.im.dendai.ac.jp*　b) *nonaka@rikkyo.ac.jp*

概要　電子書籍の利用者は増加傾向であるが，読書離れが上昇傾向となっている．電子図書館と呼ばれる青空文庫は，年間アクセス数が 1000 万件弱，作品数は 1 万 4000 件以上あるにも関わらず，作品ページに作品の内容に関する記載が一切ない．そのため，利用者はどのようなジャンルなのか，どのような内容なのかを読むまで把握できず，読書離れを促進してしまうといった問題がある．そこで我々は，小説文章を基にした自動要約システムを開発することで，自動要約された文章を利用者が読み，作品に対する興味を抱かせ，読書離れを抑えることを目的とする．事前調査として代表文抽出手法が小説作品の自動要約として有用であるかを調査し，その考察と今後の展望について述べる．

キーワード　自然言語処理，自動要約，LexRank，Indicative Summary

1　はじめに

平成 30 年度の文化庁の世論調査 [1] では，電子書籍の利用者は増加傾向である．しかし，読書量の減少や読書量を増やそうと思わない人の割合の増加から，読書離れが上昇傾向である．青空文庫 [2] のような電子図書館では 2019 年のアクセス数合計が 920 万件 [3]，作品数は文学だけでも 1 万 4000 以上 [4] ある．しかし，青空文庫で公開されている作品は要約がされていないことから，利用者は作品を読むまでどのような内容なのか把握できないという問題がある．また，どのような話が今後展開されるのか想像できず，途中で読むのをやめてしまうといったこともあり，ますます読書離れが深刻化してしまう．

そこで，我々は小説文章を基に自動要約を行うシステムの開発に取り組む．これにより，要約文章を利用者が読むことで作品に対する興味を抱かせ，読書離れを抑えることができる．青空文庫で公開されている作品には様々な作家がおり，彼らの特徴を考慮した汎用的な要約システムを作成するのは難しい．そのため，対象作品を江戸川乱歩のみとし，事前調査として LexRank[5] を用いた代表文抽出手法が小説作品の自動要約として有用であるかを検討した．

2　関連研究

小説文書の自動要約に関する研究として，Zongda ら [6] は，トピックモデリングに基づいた小説文書の抽出型自動要約を行っている．単語の分割，ストップワードの除去，ステミングなどの前処理を行い，LDA アルゴリズムを用いてトピックモデル化，候補文の重要度評価関数から機械要約を行なっている．機械要約の精度を上げるために，平滑化も行なっている．それによって，0.1%〜0.2%の圧縮率を実現し，機械要約のトピックの多様性の確保を実現している．しかし，小説の登場人物などの意味的実体を考慮しておらず，要約文に重複情報が含まれているといった問題がある．

山本ら [7] は，小説自動要約のために，SVM を用いた隣接文間の結束性判定を行なっている．小説内の指示的要約を対象とした，隣接文間の結束性の有無判定によって，文抽出手法を提案している．しかし，学習データが識別に十分反映されているとは言えず，類語を考慮していないといった問題がある．

株式会社ユーザーローカルは，芥川龍之介『蜘蛛の糸』をサンプルとした自動要約ツール [8] を開発している．このツールによって，重要と思われる文を，設定した分量で調整した結果を出力し，作品の内容が推察できるようになっている．しかし,「10 行ダイジェスト」を設定し，結果を見てみると，分量を調整するために文章の途中のみ抽出していることから，文章として不可解な部分がいくつか確認できる．また，入力文章の文字をそのまま抽出していることから，重複した単語や不可解な記号といった入力ミスも結果に反映されてしまっている．

3　提案手法

本提案手法では，形態素解析器に Juman++，抽出型要約に LexRank を用いる．日本語に対応した形態素解析器として，MeCab，Juman，Janome があるが，今回対象とする作品の制作時期が大正から昭和であるため，その時期の日本語を高い精度で形態素解析できた Juman++を採用した．

自動要約の種類として，主に抽出型と抽象型がある．前者は，対象文章から重要と思われる文を抽出して要約を作成する手法である．対象文章から文を抽出するた

Copyright is held by the author(s).
The article has been published without reviewing.

め，文法的に誤った要約にならないというメリットがあるが，文中にない単語を利用することができないため，抽象化や言い換え，接続詞の追加などができないというデメリットがある．

一方，後者は，対象文章の意味を抽象化して適切な要約を作成する手法である．対象文章にない単語を自由に使って要約を作成できるため，抽象化や言い換えが可能で圧縮率に対する自由度が高いというメリットがあるが，文法的に正しい要約の作成や隣接文間で文脈的に整合性のある文を作成することが困難であるというデメリットがある．

今回は，文法的に正しい要約を実現するために抽出型の要約を採用した．抽出型要約の中でも，グラフベースの手法，特徴ベースの手法，トピックベースの手法があるが，今回は事前調査として実装しやすかった代表的なグラフベースの手法である LexRank を採用した．

3.1 概要

青空文庫に掲載されている以下の 3 作品を対象として，前処理後 Juman++による形態素解析を行い，LexRank を用いた要約の作成を行う．

1. 江戸川乱歩『赤いカブトムシ』（青空文庫）
2. 江戸川乱歩『怪人二十面相』（青空文庫）
3. 江戸川乱歩『少年探偵団』（青空文庫）

3.2 前処理

以下の流れで前処理を行う．

処理 1. 対象作品の本文から第 3 章までの文を抽出

処理 2. （「」、。）これらの記号を半角スペースで置換

処理 3. Juman++による形態素解析

処理 4. 「名詞」・「動詞」・「副詞」・「形容詞」のみ半角スペース区切りで抽出

3.3 LexRank による要約

前処理で得られた文を圧縮率 5%，ストップワードを半角スペースとして要約を行なった.『怪人二十面相』の LexRank スコアの平均値の差を表したグラフを図 1 に示す．要約結果で情報量が少ない文を削除するために，各文の長さが 15 以下のものを削除した．3 作品の要約結果を以下に示す．

1. 江戸川乱歩『赤いカブトムシ』（青空文庫）

   三人はぞっとして、いきなりかけ出そうとしましたが、そのとき、せいようかんの方から、けたたましいさけび声がきこえてきました。まっぴるまですから、こわいことはありません。かいちゅうでんとうをてらし、長いあいだかかって、一かいと二かいのぜんぶのへやをしらべましたが、だれもいないことがわかりました。小林くんが、だんろのぼっちをおしたので、それがひらいたのです。小林くんと、だんいんの木村くんが、おばけやしきのせいようかんのちかしつで、にんげんほどもある、大きなまっかなカブトムシに出あいました。そのはらから出てきた中カブトムシも、五十センチもあるのですから、きっと作りものなのでしょう。大カブトムシとおなじことをするのです。十五センチのカブトムシのおなかに、四センチほどの四かくいあながあいて、そこから、こんどは、ほんものとおなじくらいの大きさのまっかなカブトムシが、ゆかの上にすべり出しました。

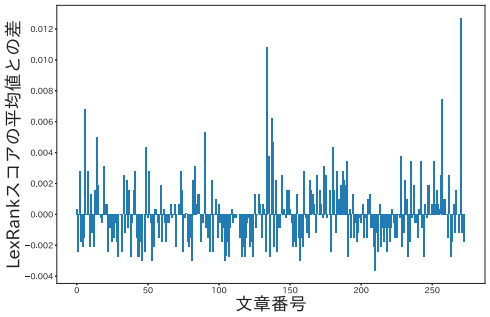


図 1 LexRank スコアの平均値との差

2. 江戸川乱歩『怪人二十面相』（青空文庫）

   その賊は二十のまったくちがった顔を持っているといわれていました。何かこれという貴重な品物をねらいますと、かならず前もって、いついく日にはそれをちょうだいに参上するという、予告状を送ることです。壮二君は、いつか一度経験した、ネズミとりをかけたときの、なんだかワクワクするような、ゆかいな気持を思いだしました。壮太郎氏と壮一君は、洋館の二階の書斎に籠城することになりました。壮太郎氏は大きな声で笑うのでした。この箱は、ここへおくことにしよう。しかし壮一君はニッコリともしません。すると、おまえは、箱だけがあって、中身のダイヤモンドがどうかしたとでもいうのか。

3. 江戸川乱歩『少年探偵団』（青空文庫）

   そいつは、暗やみの中へしか姿をあらわしませんので、何かしら、やみの中に、やみと同じ色のものが、もやもやと、うごめいていることはわかって

も、それがどんな男であるか、あるいは女であるか、おとななのか子どもなのかさえ、はっきりとはわからないのだということです。すると、ああ、どうしたというのでしょう、影はやっぱり動かないのです。かくれみのというのは、一度そのみのを身につけますと、人の姿がかき消すように見えなくなって、人中で何をしようと思うがまま、どんな悪いことをしても、とらえられる気づかいがないという、ちょうほうな魔法なのですが、黒い魔物は、それと同じように、やみのなかにとけこんで、人目をくらますことができるのでした。ところが、日がたつにつれて、お化けにもせよ、人間にもせよ、その黒いやつは、ただいたずらをしているばかりではない、何かしらおそろしい悪事をたくらんでいるにちがいないということが、だんだんわかってきたのです。まっ黒な顔の中に、白い目と白い歯とが見えるからには、こちらを向いているのにちがいありません。しかし、そんなことがあるものでしょうか。そのあいだに、怪物は女の子をつれて、どこかへ走りさってしまったのですが、では、黒い魔物は、おそろしい人さらいだったのかといいますと、べつにそうでもなかったことが、その夜ふけになってわかりました。だれでもいいから、子どもをさらおうというのではなくて、あるきまった人をねらって、つい人ちがいをしたらしく思われるのです。道化服はそんなことをいいながら、女の子の手を引いて、グングン歩いていきます。

要約文章の文字数を L_r，要約対象文章の文字数を L_s，実質の圧縮率を C_{rate} とした時の結果を表 1 に示す.

表 1 要約対象文章の実質圧縮率

作品名	L_r	L_s	C_{rate} [%]
『赤いカブトムシ』	395	5646	6.9
『怪人二十面相』	277	9866	2.8
『少年探偵団』	672	9736	6.9

4 考察

要約結果から，人物情報などの意味的情報が考慮されていないことが分かる．例えば,『赤いカブトムシ』の「三人」や『怪人二十面相』の「その賊」,『少年探偵団』の「そいつ」という単語が，要約文章を読んだだけでは誰を指しているのかを判別することができない.

また，実質圧縮率が異なることから，同じ圧縮率で要約を作成しても長い文章が選択された場合に結果が長くなってしまうという問題があることが分かる．結果が長くなってしまう問題を解決するために，ひらがなの単語を漢字に変換し，言い換えを行うなどを検討する必要があると思われる.

さらに，要約文の流れが不明瞭な点があることが分かる．例えば,『赤いカブトムシ』の 4 文目以前では「せいようかん」についての話をしていたが，5 文目以降では「カブトムシ」の話に変わっていることが分かる．このことから，要約文章の隣接文間での整合性を考慮する必要があると思われる.

最後に，登場人物がどのような人物で，どのような時代背景なのか，その他の人物とどのような関係なのかといった情報が不足していることが分かる．この問題の解決として，深層学習モデルを利用し意味的情報を取得する必要があると考えられる.

5 まとめと今後の展望

本稿では，江戸川乱歩作品を対象とした代表文抽出手法を説明し，事前調査として，LexRank を用いた小説文章の自動要約を検討した．要約結果から，代表文を抽出しただけでは，十分な要約精度とは言えないことが明らかとなった．今後の展望としては，特徴ベースの手法やトピックベースの手法を検討し，ROUGE[9] を用いた要約精度の評価を行なっていきたいと考えている．また，BERT や Transformer[10]，Reformer といった自然言語処理では一般的な深層学習のモデルを用いた自動要約手法を検討したいと考えている.

参考文献

[1] 文化庁, : 平成 30 年度「国語に関する世論調査」の結果の概要, https://www.bunka.go.jp/tokei_hakusho_shuppan/tokeichosa/kokugo_yoronchosa/pdf/r1393038_02.pdf, 2020.

[2] 青空文庫, : インターネットの電子図書館、青空文庫へようこそ。, https://www.aozora.gr.jp/, 2020.

[3] 青空文庫, : 2018 年-2019 年の年間アクセス増率分析, https://www.aozora.gr.jp/aozorablog/?p=4342, 2020.

[4] 青空文庫, : 分野別リスト, http://yozora.main.jp/, 2020.

[5] Erkan, G.,and Radev, D. R., : LexRank Graph-based Lexical Centrality as Salience in Text Summarization, Journal of Artificial Intelligence Research, Vol. 22, pp. 457-479, 2004.

[6] Wu, Z., Lei, L., Li, G., et al. : A Topic Modeling based Approach to Novel Document Automatic Summarization, Expert Systems with Applications, Vol. 84, pp. 12-23, 2017.

[7] 山本悠二, 増山繁, 酒井浩之, : 小説自動要約のための隣接文間の結束性判定手法, 言語処理学会第 12 回年次報告, 2006.

[8] 株式会社ユーザーローカル, : 自動要約ツール, https://text-summary.userlocal.jp/, 2020.

[9] ROUGE, : ROUGE, https://en.wikipedia.org/wiki/ROUGE_(metric), 2020.

[10] Ashish, V., Shazeer, N., Parmar, N., et al.: Attention Is All You Need, Advances in neural information processing systems, pp. 5998-6008, 2017.

ARG WI2 No.16, 2020
WI2-2020-09

応対履歴におけるQA関連付け手法の考察

蔦ヶ谷 文月[a]　土田 正士[b]　石川 博[c]

東京都立大学システムデザイン学部情報科学科

a) tsutagaya-fuzuki@ed.tmu.ac.jp b) tsuchida@tmu.ac.jp c) ishikawa-hiroshi@tmu.ac.jp

概要　本稿では，コールセンタの応対履歴や[1] yahoo 知恵袋のやりとりなどから，問い合わせとその回答になる文の適切な結び付けを行う手法の考察を述べる．上記のような人間同士のやりとりには記述の際の決まった型やタグのような規則がないため，それらを使った機械的な文の関連付けができない．また関連付けた文同士が適切な QA か判断する指標が一般化されておらず，人の目が必要となってしまっている．そこで TF-IDF の考え方を用いて各単語がどれくらいその文を特徴付けているかをスコア化した上で，共通単語をもつ文同士を QA 候補とした．さらにその候補から最適なものを選択するための指標として，共通単語のスコアから算出した QA スコアを用いて比較を行った．問い合わせ内容と回答が事前に結び付けられていることによって，類似内容の過去事例を参照して回答を探す際の手間を省くことができると考えられる．

キーワード　関連付け, QA スコア, TF-IDF

1.　はじめに

コールセンタでのオペレータの抱える課題のひとつに,「過去事例から類似する問題を探すのに苦労する」といったことがある. 実際, コールセンタに保管されている膨大な量の過去事例データの中から目的に合った事例を探すのにはかなりの時間を要する上, いくつかの異なる事例が候補に挙がった場合, 最適な事例を選ぶにはそのデータに対する専門的な知識が必要とされる. オペレータは専門家ではないため, 専門家に相談をしながら質問に対する回答を作成することになるが, この場合, オペレータの時間だけでなく, 相談を受ける専門家の時間も費やされるため, コスト面においてもこの課題は企業にとって非常に解消したいものだと言える.

そこで, 受けた質問に類似する内容の事例を自動で検索し, オペレータの業務の効率化ができるツールの開発をし, この課題を解決するという考えが浮かび始めた.

しかし, このようなツールが未だに実用化されていない背景として, コールセンタにおける応対履歴のような, 人間同士のやりとりがそのまま残されている生のデータには, 決まった記述の型や明確なタグ分けはなく, 質問に対する回答を適切に探し出すことが現状困難だという問題がある.

本稿では, 20 年以上にわたり IT 分野のサポートサービスを提供しており, 200 人以上のオペレータが 100 種程度の製品を扱っているコールセンタでの実際の応対履歴を分析対象とし, 自動で QA 検索を行えるツールの開発の基礎となる, 文と文の関連付け技術の問題点とその改善法について検討する.

2.　関連研究

カテゴリ分類や記述規則の無い文の集まりから QA を生成する手法として, 共通する単語で Q と A の結び付けを行う研究は過去に多くの人が試みている[1]. 出現する単語の類似度による文の結び付けは, 意味的な関連度を計ることに弱く, この手法単体では関連付けとしては弱いとされている.

川端ら[2], 奥野ら[4]は, 出現単語の表記上の近さを計算する方法は, 同義語のような意味的に同じだが表記が異なる表現に弱いとし, [2]word2vec を用いて単語の意味を拡張し, 意味的な類似度を考慮した文の結び付けを行った.

また, 吉田ら[3]は, 出現する単語の表面上の近さに加え, 回答文の記述は対応する質問文よりも狭義的な内容が示されているという QA 文の特徴を利用した, 単語の意味を予測する関連付けを行った.

上記の研究はいずれも, 表記上の単語の情報だけでは関連付けの材料としては弱いとし, 新たな情報を付加することで問題を解決しようとしたものである.

これに対して本稿では, あえて単語に同義語情報のような付加情報を与えることなく出現する単語の距離で結び付けを行い, その際に現れる問題の原因を分類し, 原因ごとの必要な対策をまとめることで, 今後の QA 生成の研究における手がかりを示した.

Copyright is held by the author(s).
The article has been published without reviewing.

[1]https://chiebukuro.yahoo.co.jp/

[2]https://code.google.com/archive/p/word2vec/

3. 分析手法

本稿で扱う応対履歴データは，応対履歴が件ごとのファイルにテキストデータとして保存されている．1 文対 1 文の QA ペアを作るにあたり，それぞれのファイル対して，以下の手順を施し分析の準備をする．

1. Q 文と A 文の抽出
   読点（。），記号（？）を文末の判定とし，文を抽出する．さらに各文に含まれる文末表現を用いて，質問文（Q 文）・回答文（A 文）の文体の規則を評価して該当する文が質問文であるか回答文であるか，該当しないかを決定し，ファイルごとの Q 文，A 文リストを作成する．

2. 同一文（繰り返し部分）の除去
   Q 文，A 文に対して Python の形態素解析エンジン「[3]janome」を使用し，文を単語単位に分解する．このとき，内容文は同一でも文頭に件名やラベル記述があり，除去したい文を除去しきれない場合を考慮し， 8 割以上の単語がファイル内の別の文と一致する Q 文，A 文に関しては，同一文の繰り返しと判定し，先に出現した文のみをリストに残す処理を行う．

3. 単語の抽出
   手順 2. で分解した単語から名詞のみを抽出し文ごとの名詞リストを作成する．このとき，英数同士，漢字同士，カタカナ同士で名詞が連続している場合，連続している名詞をまとめて連接語とみなし，名詞リストに連接語を加える．また，記号，数字のみで構成された名詞や文字数が 1 文字の名詞は経験的に文の特徴付けにあまり役立たないと判断し，stop word として名詞リストから除外した．

4. 単語の特徴量を算出
   手順 4. で得られた名詞リストに対して，Q 文，A 文それぞれ別のベクトル空間で TF-IDF 法を用い，各文における単語の特徴量$weight_{t,d}$を以下の式より算出する．

$$weight_{t,s} = tf_{t,s} * \log_2 \frac{N}{sentence\_freq_t} \quad (1)$$

$tf_{t,s}$は任意の単語 t の文 s 内での出現回数，$sentence_freq_t$は単語 t が出現する文の数，N は同じベクトル空間内の文の総数を示す．このとき，単語の特徴量$weight_{t,s}$を算出する際，自然言語処理ライブラリの「[4]gensim」を用いた．（gensim の仕様で特徴量$weight_{t,s}$は正規化されている）

5. Q 文と A 文の結び付け
   共通の名詞をもつ Q 文と A 文を結び付け QA ペアを作る．このとき，手順 5.で求めた名詞ごとの特徴量$weight_{t,d}$に応じて以下の式より QA ペアのスコアQA_{score}を算出する．

$$QA_{score} = \sum \sqrt{weight_{t,Q} * weight_{t,A}} \quad (2)$$

$weight_{t,Q}, weight_{t,A}$はそれぞれ共通名詞 t の Q 文，A 文での特徴量を示す．

4. 評価方法

「QA_{score}の高いペアは，質問に対して適切な回答が結びついている」と仮定し，以下の手順で評価を行う．

1. 生成された全 QA ペア（13359 件）のうち，一つの Q に対して最もスコアの高いペア（1318 件）を抽出し，これらを各 Q に対する適切なペアであると仮定する．それらをスコア順にソートする．

2. 1. でソートしたデータの内の上位 30 件，中央値付近の 30 件の QA ペアに関して，専門家 3 名と著者とで QA の組み合わせとして適当と思われるものに○，そうでないものに×を付け，人間の観点での QA としての妥当性を見て，QA スコアの優劣で何かしらの傾向が顕在化してはいないか探る．

3. ×と判定された QA ペアの原因を，スコアの観点，名詞抽出の観点，文の切り方の観点から以下の 5 つのパターンに分類する．
   ① A 文に，Q 文と同一の文字列（url，メールアドレス，パス，質問の繰り返し）が含まれており，その部分に共通語が多く含まれスコアが高くなっている場合．
   ② 文の区切り方やタグ付けの段階での問題の場合．
   ③ 同一文字列ではないが，マニュアルや過去事例，内部のやりとりの引用が Q 文や A 文に含まれていることで，共通語が相対的に多くなり，スコアが高くなっている場合．
   ④ 原因①や③とは逆に文全体に名詞が少ない，また，そのペアでしか出現しない共通語があり，1 名詞あたりの特徴量が高くなったことで，相対的にスコアが高い場合．

[3]https://mocobeta.github.io/janome/
[4]https://radimrehurek.com/gensim/

⑤ Q 文が抽象的すぎる，元の応対履歴を確認しても該当する A が存在しない，回答が複数文にまたがっているなどの理由で，共通語の結び付けでは適切な1文対1文の QA ペアを作ることが困難な場合．

5. 考察

表 1 から，スコア上位のペアは中央値付近より QA として適切と判定された割合が高くなっており，QA スコアでの適切さの判定がある程度有効であることは確かめられたものの，上位 30 件でも約半数のペアは QA として適切でないものが入ってしまっているとも取ることができる．また，表 2 に見られるように，適切でないペアの原因は，QA スコアによって偏りがあることがわかる．以下この章では，不適切なペアのスコアが高い，適切なペアのスコアが低いことへの対策を，原因ごとに考察していく．

表 1：適切だと判断した QA ペアの数のまとめ

	A	B	C	著者
上位 30 件	18	13	16	12
中央値付近 30 件	9	6	5	4

表 2：課題の種類の分布（重複あり）

	上位 30 件	中央値付近 30 件
原因①	11	0
原因②	7	8
原因③	3	2
原因④	1	4
原因⑤	4	15

① 上位 30 件では原因①による不適切なペアが最も多いが，これはQA_{score} を求める式(2)が，共通語の数の和を取っているため，共通語が多く取れるとスコアが高くなるためである．図 1 は，上位 30 件における適切と判断されたペア（左）と，原因①に分類された×判定のペア（右）の 1 共通語あたりの特徴量の分布を示したものである．これを見ると，原因①に分類されたペアは 1 共通語あたりの特徴量が低い傾向にあることが分かる．実際，スコアが最高値の QA ペアは共通語の数も 55 語と全 QA ペアの中で最も多いが，1 共通語あたりの平均スコアは 0.1259 となっており，図 1 左側の適切な QA ペアの分布と比較するとかなり低いことが分かる．このことから，原因①によってスコアが上位になっている不適切なペアに関しては，1 共通語あたりの平均スコアを比較することで順位を落とすことができると考えられる．また，単純に一定長の同じ文字列を共通語判定時に取り除くという対策も考えられるが，この場合，どれくらいの長さの文字列を同じと扱うのか，また，対象とする同じ文字列が内容的に文中から取り除いてもよいものかどうかの判定を行う必要がある．

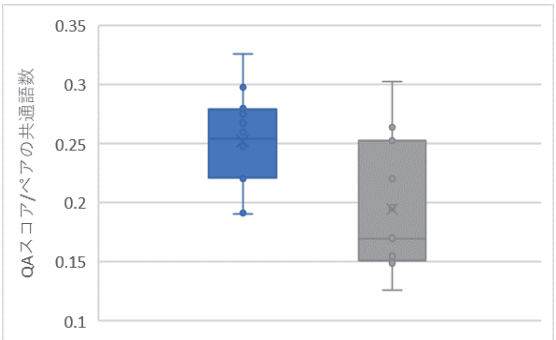


図 1： 1 共通語あたりの特徴量の分布

② 原因②は上位 30 件，中央値付近 30 件のどちらともで一定数の例があるが，その原因をさらに細分化すると以下の表にまとめることができた．

表 3：原因②の細分化まとめ

	イ)	ロ)	ハ)
上位 30 件	0 件	7 件	0 件
中央値付近 30 件	5 件	0 件	3 件

イ) その他のラベルに分類された文に回答候補が存在した．

ロ) 「。」「？」が無いために文が区切れず，2 文以上が 1 つの文とみなされた．

ハ) 文の流れでは区切る必要のないところに「。」「？」があり，文を区切りすぎてしまった．

イ）は，その他のラベル文に回答になりうる文があった場合なので，その文に一定数の共通語があれば，語尾以外の判定基準を見つけラベル付け精度の改善をすることで適切な QA ペアを作ることができると考えられる．ロ），ハ）に関しては，文を区切る条件に新たな基準を設けることで改善することができる．

③ 原因③にあてはまる QA ペアも原因①と同様に不要な共通語が多く取れてしまっていることが要因の一つとなっている．そのため，1 共通語あたりのスコアで比較を行い適切な QA ペアを得るという対策も取れるが，他にも，3. 分析手法の 3.「単語の抽出」で示した通り，本稿では「確認方法」を「確認　方法　確認方法」のように，本来 1 語でとられるべき名詞を 3 語と取っているので，この場合を「確認方法」の 1 語のみを取るようにすることで，

上位の不適切なペアから不必要な共通語を減らすことができる. しかし, 連接前の単語を結び付けに必要としているペアも中には存在するため, 連接前の単語を完全に除去するのではなく, 特徴量に重み付をして連接語と連接前の単語の重みに差をつけるといった対策が有効と考えられる. また, 下記の図 2 に示すように,「B とありますが, B は A ですか？⇒A です」と結び付けたいにもかかわらず, B の部分に特徴的な名詞を多く含むマニュアルや内部者同士のやり取りが含まれている場合が多く, A 側よりも B 側に QA スコアが引っ張られてしまうという事例も原因③が起こる要因として確認されている. この場合,「, 」などで区切れている部分などを指標に, 文の部分ごとに名詞の特徴量の重みを変えるという文法的アプローチで A の内容の回答を結び付けることができるようになる.

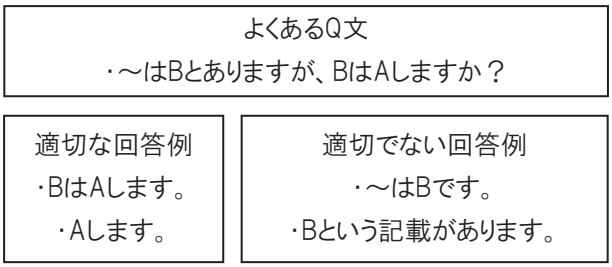


図 2: QA によくある Q 文とその回答パターンの例

④ 原因④への対策としては, 適切と思われる A 文の名詞リストから「こと ため」のような余計な名詞を取り除くための stop-word-list を作成し, 1 名詞あたりの特徴量を高くするという対策が考えられるほか, 原因③と同様に, 名詞の出現する位置によって特徴量に差をつけることで, 相対的に小さくなっている適切な QA ペアの 1 共通語あたりの特徴量を大きくすることで解決ができる.

⑤ 原因⑤は大まかに,「Q 文が現段階では QA ペアを作るのに相応しくない文であった」という括りである. 原因⑤にあてはまる事例の1つに,「Q 文が抽象的で QA の質問として適切でない」という事例がみられたが, これは質問の根幹になりうる部分を,「その」「先ほどの」「以下の」などの指示語を使って別の文から引用しているため, その1文だけ見ても何の話なのかが分かりづらくなってしまっているからであった. 対策としては, 応対履歴から文の切り出しをする際に, 文の出現する順番などから, 話者ごとの文のまとまりを作り, 1文対1文のペアだけでなく, 場合によって複数文対複数文のペアも作れるようにする, 1文対1文のペアで結び付けを行ってから, 各文を1人の話者が話している範囲まで拡張するといった案が挙げられる. また, この対策は,「Q 文が複数文にまたがっていてペアが作れない」という原因に対しても有効である.

6. おわりに

表記上の単語での文の結び付けがうまくいかない原因に対する改善手法と, その際に必要となることを図 3 にまとめた.

分析手順1, 2, 3で改善
原因① ⇒特定の文字列の除去
・抜き取る文字列を判定する指標の選定
原因② ⇒文の区切り方の工夫 ・ラベルの振り方の工夫
・文を区切る基準とラベル付けの指標の改良
原因④ ⇒stop-word-listの作成
・余計な語彙である判定をする指標の選定
分析手順4, 5で改善
原因③ ⇒引用文に含まれる名詞の特徴量に重み付をする
・妥当な重み付の指標の選定
原因③④ ⇒文法的アプローチで特徴量の重み付けをする
・どのあたりまでの文法に言及するかの検討
QAペアを作ってから行う改善
原因①③ ⇒QAスコア以外の指標でペアの適切さを判定
・適用する場合の具体的な指標の選定
原因⑤ ⇒周辺文まで拡張して適切なQAにする
・拡張する範囲や基準の具体化
原因⑤ ⇒候補に1つも適切なペアがないQ文を排除
・適切でないかどうかの判定を見分ける基準の具体化

図 3:各原因に対する対策と必要なことのまとめ

謝辞

本研究の一部は, JSPS 科研費 20K12081, 野村マネジメントスクール研究助成及び東京都立大学傾斜的研究費(全学分)学長裁量枠国際研究環支援による.

参考文献

[1] 栗山和子, 神門典子:Q&A サイトにおける質問と回答の分析, 情報処理学会研究報告, pp. 3-7,, 2009.

[2] 川端貴幸, 佐藤一誠:意味と表記の組み合わせによる用例ベースの質問応答モデル, The 31st Annual Conference of the Japanese Society for Artificial Intelligence, pp.1-4, 2017.

[3] 吉田知訓, 間瀬心博, 北村泰彦:質問応答 Web サイトからの関連語ネットワークの自動抽出, 社団法人 電子情報通信学会, pp.1-6, 2008.

[4] 奥野翔太, 荒木健治:単語の分散表現により獲得した類義語を用いた FAQ 検索システムの評価性能, WI2-2017-6, pp.23-24, 2017.

ARG WI2 No.16,2020
WI2-2020-10

人間の情動理解のための感情生成モデルの構築手法検討および生成自動化手法の模索

茂島 祐太† 當間 愛晃‡

†琉球大学大学院理工学研究科情報工学専攻

‡琉球大学工学部工学科知能情報コース

†k198571@ie.u-ryukyu.ac.jp ‡tnal@ie.u-ryukyu.ac.jp

概要 人には心と呼ばれるものがある. 本研究では心についての理解を深めるために, 心を情動と意識の二つに分け, さらに段階的な理解のために情動のうち, 感情を生成するプロセスを観察することを目的とする. そのため, 現在感情の生成において提唱されている, 末梢起源説および中枢起源説の二つについてモデル構築を行い, 表現可能数を比較することにより計算機を用いた感情生成の表現により適した説を検討した. 結果としてわずかに末梢起源説の方が表現可能数が多かった. しかしながら, より表現可能数を増やすべくデータセットのバリエーションの向上や, モデルの複雑化を目指す.

キーワード 感情, ユーザーモデル

1 研究背景と目的

人間には心と呼ばれる感覚器が存在する. これは特権的で通常他の人間との本質的な共有は行うことができず, ある種のブラックボックスとなっている. そこで本研究では心の理解を深めるために情報工学の分野からのアプローチを図ることを主目的とする. その足がかりとして心の中でも特に情動, さらにその中の感情に焦点を絞る.

この感情を理解するためにまず, 感情生成を再現したモデルを作成, 観察することを目標とする. そういったモデルを作成することで, 実際の人間には与えられないような状況下での感情発生や, あるいは単に観察そのものが透明化され行いやすくなるものと推察する.

本原稿では, 計算機を用いたモデル作成により適した理論を模索するため, 現在心理学の分野で唱えられている中枢起源説と末梢起源説を簡易的なモデルに実装し, 比較を行うものとする.

2 前提知識と定義付け

2.1 心について

心とは主に生物に備わった本人のみがアクセスしえる特権的な器官のことである. 言葉や表現で伝えることはできても, 本質的な理解を行うことは甚だ難しく, ほとんど不可能だと言える. この心は, さらに意識と情動に分けることができると考えられる.

意識とは, 本人が自覚的に操作し得る部分であり, 思考などの知的な活動を担う部分である. 基本的には言語化が容易で, 比較的共有しやすい.

情動とは, 心の中の無意識的な部分であり, さらに感情(情動感覚)と反射的な行動(情動行動)に分けられる. 意識的に情動を制御することは難しく, また他者との共有も難しいと考えられる.

2.2 感情について

人間の心の中でも, 特に情動に含まれる感覚のことである. 非常に特権的であり, 自身が感じている感情を他者に本質的な意味で共有することはできず, また離散的ではなく連続した感覚であるため一様に言語化するのが非常に難しい. これに対して一定の尺度と解釈を与えたのがRobert Plutchikの提唱した感情の輪である(図1).

Plutchikによると, 感情には8つの基本的感情が存在し, このほかの様々な情動は各基本的情動の合成である. また, それぞれ輪の対極に位置する感情同士で対になっていると考えられる. そのため, この輪の通りに感情が連続的に変化するとは限らないことに注意が必要である[1].

本研究で取り扱う感情とはPlutchikの感情の輪で示された8つの基本的感情, すなわち喜び, 信頼, 心配, 驚き, 悲しみ, 嫌悪, 怒り, 期待の8つの感情のこととする.

2.3 末梢起源説について

末梢起源説とは, William Jamesによって1894年に提唱された情動理論である. これに先駆けてCarl Langeによって主張されていた内容を包括するため, ジェームズ

Copyright is held by the author(s).
The article has been published without reviewing.

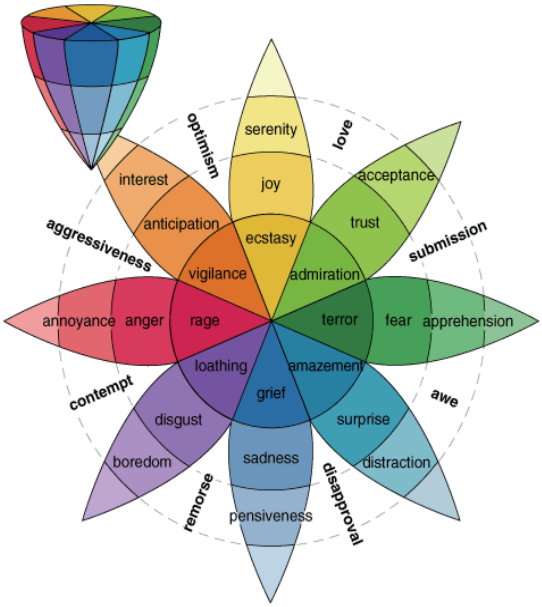


図 1 Plutchik の感情の輪

出典：American Scientist Vol.89, No.4 P.349

=ランゲ説とも呼ばれる.

この説の主張するところによると，人間の感情は外部からの刺激によって引き起こされた身体的行動や反応が脳に伝達されることによって初めて感情が引き起こされるということで，端的に言えば「怒るから殴る」のではなく「殴るから怒る」ということになる.

この反応というのは身体表層部分，つまり表情や動作だけではなく内臓等の分泌なども密接に関わっており，表層的な行動のみを真似たとしても感情が起こりえないという[2].

またこの説は，後述する中枢起源説における実証実験により問題点が指摘され，一時期淘汰されたかのように見えたが，近年ではSilvan Tomkinsにより提唱された表情フィードバック仮説などにより，一部再評価されている[3].

本研究においては，データ収集を一般的なアンケートによってのみ行ったため，具体的な内分泌系のデータを揃えることができない．そのため，アンケート内容として該当の感情が起きた時に意識的に行ったわけではない，とっさに取った行動，あるいはとっさに出た反応を思い起こしてもらい，それを末梢起源説における行動や反応として定義している．例としては，怒りの感情が起きた際に顔が熱くなった，喜びの感情の際に口元が緩んだ，驚きの感情の際にうわっと言った，等である.

2.4 中枢起源説について

中枢起源説とは，W. B. Cannonによって1927年に提唱された情動理論である．その後Philip Birdによって仮説の実証が行われたため，二人の名を取ってキャノン=バード説とも言われる.

この情動理論は前述した末梢起源説に対する批判として唱えられたものであり，人間の感情は外部刺激を感覚器が受け取ると，インパルスが脳に向けて発射され，これを皮質，および視床下部が受け取ることにより脳によって刺激が評価され，該当する感情が呼び起こされるものとされている．これは我々のもっとも素朴な理解に近しいものであると言える.

また，前述の末梢起源説に対しての批判として動物実験を行い生存可能な程度内臓を切除した動物も普段と変わらぬ感情を見せたこと，逆に薬物等を用いて内臓の分泌のコントロールを行なっても同様の感情が起こらなかったことをあげ，内臓の分泌が必ずしも感情を発生させるのに必要ではないということをあげた.

またキャノンは，ジェームズらが説明していない言葉にできない衝動についてもまた説明を与えている．視床の細胞がなんらかの原因で皮質と連動せずに発火した場合，具体的な認知を伴わないまま不明瞭な感情が呼び起こされることがあり，これを衝動であるとしている[4].

本研究では，外部刺激による感覚器の興奮をすなわち感情が発生した時の周囲の状況であると定義している．例としては，汚物を見た時に嫌悪の感情が起きた，叩かれた時に怒りの感情が起きた，病気の母がいた時に心配の感情が起きた，等である.

2.5 Marrの3つの水準について

Marrの3つの水準とは，David Marrが1982年に提唱した，複雑なシステムを理解するためには3つの水準に分ける必要があるという計算論的神経科学の考え方である.

Marrによると，3つの水準とはすなわち計算理論の水準・アルゴリズムの水準・インプリメンテーションの水準の3つに分けられる．より具体的には，計算理論の水準とは一体何が目的なのか，そこにはどのような制約があるのかと言った視点から考察をする．アルゴリズムの水準とは，前述の計算理論の水準で設定された目標と制約を満足するためにどのような処理を行うのか，その入力と出力はなんなのかを考察する．インプリメンテーションの水準とは，実際の物理的挙動はどうなっているのか？という視点において考察することである[5].

3 本研究の狙い

2節にあげた前提知識と定義を踏まえ，改めて本研究の狙いについて説明する.

本研究では，度々あげている通り心を特権的で他者

に閲覧不可能であるものとして捉え，情報工学的アプローチから段階的に理解を深めていくために心理学において唱えられている二つの説の実装を通して感情の生成モデルを計算機上で実装していくことを目的とする．

計算機上で実装した両者のモデルのうち，より現実に即した表現を簡易に行える説を採択することにより，さらにモデルを洗練，拡張することが可能にすることが本研究の狙いである．

4 実験

4.1 実験概要

本実験では，計算機上に中枢起源説と末梢起源説に対応する多次元辞書を持ったモデルを作成し，そのモデルに対して辞書内の単語を網羅的に組み合わせたデータセットを読み込ませることにより擬似的に多様な状況，あるいは行動や反応を作る．それに対して応答できなかった組み合わせの数をカウントすることによりどちらの方がより多くのパターンに対して応答できるかを測る．

4.2 事前準備

あらかじめアンケート調査を行い，基本8感情が発生した時の状況・行動や反応・その時起こった感情の種類について文章で回答してもらった(図2)．得た回答は各8感情について怒り以外の感情についての回答が5件ずつ，怒りの感情についての回答が6件，総回答件数は41件であった．

文章で得た回答から，状況・行動それぞれについて動詞および名詞とそれを修飾している語句のセットを抜き出し，感情タグと組み合わせることによってデータセットを作成した(表1)．1行が1つのアンケート結果になっており，1列目が状況の動詞および名詞，2列目が状況の修飾語，3列目が行動および反応の動詞もしくは名詞，4列目が行動および反応の修飾語，5列目が感情タグである．2列目および4列目の修飾語に関しては，特に修飾している語句がなければ，なしと記載してある．

4.3 モデルの作成

モデルには多次元辞書を用いて特定の動詞または名詞に対し，それを修飾する語句によって出力感情が変化するように構築した(図3，4)．中枢起源説の場合，何かを見たという外部からの刺激があった場合，まず辞書の中の“見た”というキーを検索する．ヒットした場合，一体何を見たのかをデータセットに記載されている修飾語で検索し，登録されている値である感情を出力する．同様に，末梢起源説の場合，熱くなるという反応があった場合，辞書の中で”熱くなる”というキーを検索する．ヒットした場合，何が熱くなったのかをデータセットに記載されている修飾語で検索し，登録されている値である感情を出力する．

また辞書内に存在しない組み合わせ，すなわち出力すべき感情がない時は該当なし，一通りの組み合わせに対して出力すべき感情が複数ある場合は複数該当ありとし，具体的な感情を出力しないようにした．

表 1 データセットの一部抜粋

状況の動詞/名詞	状況の修飾語	行動/反応の動詞/名詞	行動/反応の修飾語	感情タグ
出てきた	ゴキブリ	言う	うわっ	驚き
出てきた	虫	払う	手	驚き
腐った	食べ物	捨てる	なし	悲しみ

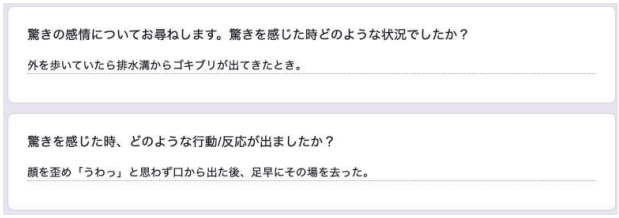


図 2 アンケート結果の一部抜粋

```
scene = {
    "叩かれた":{"なし":"怒り"},
"攻撃された":{"なし":"悲しみ"},
"聞いた":{"説明":"信頼"},
"目の当たりにした":{"なし":"驚き"},
"見た":{"プレゼント":"期待","嫌いな人":"嫌悪","汚いもの":"嫌悪","なし":"喜び"},
"来た":{"迎え":"信頼"},
"行った":{"見知らぬ土地":"心配"},
"妊娠していた":{"友人":"驚き"},
"居た":{"子供":"悲しみ"},
"言われた":{"文句":"嫌悪"},
"受けた":{"診断":"怒り"},
"応募した":{"懸賞企画":"期待"},
```

図 3 中枢起源説モデルの多次元辞書(抜粋)

```
scene = {
    "熱くなる":{"顔":"怒り","目頭":"喜び"},
    "出る":{"よだれ":"喜び","涙":"嫌悪"},
    "つく":{"傷":"悲しみ"},
    "真っ白になる":{"頭":"驚き"},
    "緩む":{"口元":["期待","喜び"]},
    "歪む":{"口元":["嫌悪","心配"]},
```

図 4 末梢起源説モデルの多次元辞書(抜粋)

出る	頭	該当なし
出る	口元	該当なし
出る	口元	該当なし
出る	鼓動	該当なし

図 5 出力結果の一部抜粋

4.4 実験内容

実験はデータセットを読み込んだのち，状況の動詞や名詞とそれを修飾する語句を網羅的に組み合わせ，中枢起源説モデルに対して入力する．出力された結果のうち該当なしの数および複数該当ありの数をカウントし，様々な状況に対する感情生成状況を評価する．

表 2 各モデルの感情生成結果

組み合わせ	中枢起源説[件]	末梢起源説[件]
総件数	1681	1681
該当なし	1604	1575
複数該当あり	0	26
説明可能	77	80

同様に，行動および反応の動詞や名詞とそれを修飾する語句を網羅的に組み合わせ，末梢起源説モデルに対して入力する．出力された結果のうち該当なしの数および複数該当ありの数をカウントし，様々な状況に対する感情生成状況を評価する．

4.5 実験結果

各モデルの感情生成結果を挙げる(表2)．

ただし，これらの組み合わせは複数の動詞および名詞にまたがって重複する修飾語が用いられている場合でも，考慮せずにそのまま組み合わせているため何度か同じ組み合わせが出てきている(図5)．

動詞および名詞の数，修飾語句の数は各モデルともにそれぞれ41個ずつであった．

4.6 考察

以上の結果を踏まえて本実験の考察を行う．

単純に，表現できた組み合わせの数を鑑みると，末梢起源説モデルのほうが生成されなかった数は少ないため，末梢起源説の方が計算機上で感情生成モデルを表現するのには有利であるように読み取れる．しかしながら，両モデルとも生成できた数は生成できなかった数に比べて少数であり，かつ出力すべき感情が複数ある場合について対応できていないため必ずしも優れているとは言えない．また，データセットの量も少なく，アンケート結果からの単語の抜き出しをある程度恣意的に行なっている．例えば，図2のアンケート結果から単語を抜き出したものが表1のデータセットとなっているが，アンケート結果中に含まれる「顔を歪め」や「足早にその場を去った」などは抜き出していない．これは，一つのアンケート結果から複数の状況や行動，反応を抜き出してしまうと，各感情のデータセットの数に偏りが発生してしまい，好ましくないと考えたためである．

そのため，単語の偏りが発生している現時点ではどちらがよいか判断しかねる．

5 まとめと今後の課題

心についての理解，情動についての理解を深めるため心理学の分野において感情生成について提唱している末梢起源説，中枢起源説の二つの説をMarrの3つの水準に基づいて計算機上で簡易的なモデルとして実装し，感情生成数のカウントを行い感情生成状況の評価をした．結果，わずかながら末梢起源説モデルのほうが良い成果を出したものの，データセットや単語選択，モデルそのものが未熟でありこの結果を持って末梢起源説の方が優れているとは言えないのが現状である．

今後の課題として，より大規模なデータセットの構築，およびアンケート文章に対してより詳細に対応するためにモデルの拡張が必要になると考えている．また，単語の抜き出しが恣意的であるため，今後は自動化手法として要約を中心とした自然言語処理的アプローチにも着手していきたい．

参考文献

[1]宇津木成介：基本的感情の数について，国際文化学研究：神戸大学大学院国際文化学研究科紀，Vol.29，pp.73-91，2007．

[2]James，W著，今田寛訳：心理学（下），pp.201-224，1993

[3]守秀子：「笑う門には福来る」表情フィードバック仮説とその実験的検証，文化学園長野専門学校研究紀要，Vol.5，2013．

[4]Cannon.W.B著，宇津木成介訳：ジェームズ・ランゲの情動理論：その検証と代替理論，近代，Vol.100，pp.43-70，2008

[5]国里愛彦，片平健太郎，沖村宰，山下祐一：計算論的精神医学，pp.14-17，2019．

ARG WI2 No.16, 2020
WI2-2020-11

推薦システムにおける推薦者のアイテム受容に与える影響に関する基礎調査

松嶋 理香子† 土方 嘉徳† Shlomo Berkovsky‡

†関西学院大学商学部

‡Department of Computing, Macquarie University

contact@soc-research.org

概要 推薦システム(RS)は幅広く活用されているが，多くのユーザは他のユーザの口コミも参考にしており，特に実世界の友人や専門家からの口コミは，より強い影響力を持つと言われる．本稿では，推薦結果の提示において，推薦者を提示することがユーザのアイテム受容に影響を与えるか否かを調査した．また，推薦者として実世界で親しい「友人」と，専門家の「映画コメンテーター」，より人間らしい RS である「映画ロボット」，既存の RS と同じ「推薦者なし」の間でアイテム受容への影響に差があるかを調べた．被験者実験の結果，推薦者の有無は受容に影響しないが，推薦者の種類によって受容に差があり，特に「友人」はより強い影響を持つとわかった．

キーワード 推薦システム，インタフェース，ソーシャルインタラクション，アイテム受容

1 はじめに

推薦システム(RS)についての研究は，1990 年代初期に行われ始め，当初は推薦の精度を高めることが目的とされた[1-3]．しかし，2000 年代に，推薦精度が必ずしもユーザの満足を満たさず，RS 全体の使い勝手の良さが改善されるべきだと主張されるようになった[4]．近年は，ユーザ経験を改善するための研究が行われている[5,6]．

RS の処理の過程は，O-I-P モデル (Output – Input - Process model)[7]で表されることが多い．このモデルでは，ユーザがアイテムへの評価値を入力するインプットの段階と，ユーザの好みを予測するプロセスの段階，推薦結果を表示するアウトプットの段階の 3 つに分けられる．これらの段階のすべてにおいて，使い勝手が良くなるほど，RS におけるユーザ経験は改善されると思われる．

本研究では，アウトプットの段階に注目する．現代のユーザは RS と関わる機会を多く持つが O-I-P モデルの 3 段階の中で，ユーザが最も RS を意識するのは，ユーザにお薦めが表示されるアウトプットの段階であると思われる．そのため，そのインタフェースはユーザの購買または選択の意思決定に影響を持つと考えられる．

これまで RS のインタフェースについて多くの研究が行われている．中でも Staffa と Rossi の研究[8]は，スマートフォンでの一般的な映画推薦とヒューマノイドロボットでの映画推薦を，ユーザのアイテム受容（推薦されたアイテムを受け入れて消費するかどうか）の点で比較している．実験の結果，2 つの出力方法においてユーザのアイテム受容に統計的な差はないが，スマートフォンより，人間味のあるヒューマノイドロボットでのインタラクションは魅力的で経験の効果を高めるとわかった．また，Bonhard と Sasse の研究[9]では，RS によるお薦めの信頼性をユーザが簡単に判断できるようにするデザインを検討した．個人の嗜好が伴う分野(音楽・映画など)では，多くの経験を持つ人や，実在する人物による説明がユーザにとって有益だと報告された．

第 3 者からの口コミは購買意思決定に大きく影響すると言われており，オンライン上の口コミを読んだユーザの 90%は，自身の意思決定に影響を受けたという調査結果もある[10]．また，Gellerstedt と Arvemo の研究 [11]では，ホテルに関するオンライン上のレビューと，実世界の友人のレビューの影響を比較している．その結果，友人からの否定的なレビューはオンライン上の過半数の肯定的なレビューの影響力を上回り，また，友人からの肯定的なレビューはオンライン上の過半数の否定的なレビューの影響力をも上回るとわかった．ZhouとDuanの研究[12]では，ソフトウェア市場サイトでのオンラインユーザと専門家のレビューの影響に注目している．CNET のダウンロードサイト上に存在するレビュー数やダウンロード数を分析した結果，専門家のレビュー数は，オンラインユーザの製品選択とレビュー数に正の相関があるとわかった．

本研究では，RS が出力する推薦結果において，推薦者を提示することのアイテム受容に対する影響の有無を明らかにする．本稿では，ユーザが推薦アイテムを

Copyright is held by the author(s).
The article has been published without reviewing.

購買/選択することを「受容」と呼ぶ. 推薦者を提示するインタフェースと提示しないインタフェースを作成し, これらのインタフェース上で被験者実験を行うことで, 検証する. 実験で用いる推薦のドメインは, 推薦システムの研究では最もよく用いられる映画とする. 本研究の貢献は以下の通りである.

- RS における推薦者表示の有無が, ユーザのアイテム受容に影響があるかどうかを明らかにしたこと
- RS で推薦者を表示した場合, 友人, 専門家, ロボット, 推薦者なしのいずれの種類の推薦者が最もユーザのアイテム受容に影響を与えるかを明らかにしたこと

本稿の構成は, 以下の通りである. 2 章で実験方法について, 3 章で実験結果と考察について, 4 章でまとめを述べる.

2 実験方法

2.1 推薦者の種類

本研究で注目する推薦者の種類を説明する.

- 友人

「友人」は, 被験者の実世界での友人となる. [11]の研究で, 実世界の友人の口コミは意思決定への影響が強いことが確かめられている. そこで, 本研究でも推薦者の種類の一つとして友人を対象とする.

- 映画コメンテーター

「映画コメンテーター」は, 映画の専門家である. [9]の研究で, 専門家による説明はユーザに有益と捉えられる傾向があることが確かめられている. そこで, 本研究でも推薦者の種類の一つとして映画コメンテーターを対象とする. 実験では, 日本で活躍しており比較的年齢が若い「有村昆」(男性)と「LiLiCo」(女性)を推薦者とする.

- 映画ロボット

研究[8]で, より人間らしい RS は, ユーザ経験を高めより満足を満たすことが分かっている. そこで本研究でも, 従来の RS のインタフェースよりも, より人間味のあるインタフェースとして「映画ロボット」を対象とする. 本研究で用いる実験用の RS では,「映画ロボット」は, 知能の実体として存在するものではないが, 被験者には映画に関する知識を搭載した人工知能だと説明し, プライミングを行う.

- 推薦者なし

「推薦者なし」では, 一般的な RS のように, 推薦結果において推薦者は表示しない. これにより, 推薦者を表示しない場合よりも, 上記の 3 種類の推薦者を提示することが, どれだけ影響力を持つかを明らかにすることができる.

2.2 実験の基本方針

本研究では, 推薦者の提示の有無や推薦者の種類のアイテム受容に対する効果を測定することができるよう, 独自に実験用の推薦システムを実装する. 実験用の推薦システム(以降,「実験システム」)は, オンラインで参加できるように, Nuxt.js により Web アプリケーションとして実装した. 被験者には, PC やスマートフォンなど, 任意の端末から実験に参加できるようにした. 推薦者に「友人」があるため, メンバーが互いを認識している実際の組織単位で実験を行う. 被験者には, 実験システムのウェブサイトの URL を配布し, 被験者は各自で実験に回答する. 実験では, 被験者に映画を推薦し, 被験者は気に入った映画を購入する(ただし, 実際にお金を払うわけではない). 購入を行うことを「受容」とみなす. 実験システムでは, ユーザの操作ログを詳細に記録しており, アイテムを購入したかどうかのデータも取得している.

実験は, 被験者をグループ A と B に分け, A では推薦者ありの映画リストで, B では推薦者なしの映画リストで推薦する. 推薦者ありの実験では, 1 つの推薦リストにおいて, 4 種類の推薦者(「推薦者なし」を含む)が 2 本ずつ映画を推薦するようにする. 被験者は, この推薦リストによる推薦を 4 回受ける. 推薦者ありの映画リストにおいて, 推薦者として「友人」が表示される場合は, それは同じ組織内の他のメンバーになり, その名前とアバターが表示される.

この実験のために, 我々は独自に映画のデータを収集した. 具体的には, 1980～2019 年の国内年間興行収入ランキング[1]から映画のタイトルを収集し, それらからシリーズ物の重複を削除して, 実験に用いる映画のタイトルの候補とした. なお, ここで収集した映画は洋画に限定した. さらに, 映画データベースサイトの IMDb[2]が付与するジャンルに従い, 各ジャンルに一定数の映画タイトルが含まれるようにランダムに映画のタイトルを選択し, 実験用データセットとした. 実験用データセットには, 全部で 570 本の映画が収録されることになった. ユーザに提示する映画の情報として, 150 文字程度の長さの映画の紹介文と, 映画のパッケージの画像を映画.com[3]の Web サイトから収集した.

2.3 実験用推薦システムのアルゴリズム

実験用推薦システムで採用した推薦アルゴリズムについて説明する.

推薦アルゴリズムには, ジャンルに基づくアルゴリズムを採用する. このアルゴリズムでは, 事前にユーザに好みの映画のジャンルを尋ねておき, そのジャンルに適合する映画を推薦する. 実験で用いる映画ジャンルは, ア

[1] https://entamedata.web.fc2.com/movie/top_jmovie.html
[2] http://imdb.com
[3] https://eiga.com

クション, アドベンチャー, アニメーション, コメディ, クライム, ドラマ, ファミリー, ホラー, ヒストリー, ミュージカル・音楽, ミステリー, ロマンス, SF, スリラーの 14 種類で, IMDb を参考に設定した. アルゴリズムでは, ユーザの好みに合うジャンルごとに, 設定した本数分の映画をランダムに選択する. ユーザには, 事前に映画ジャンルの好みを 0(全く好みでない)～3(非常に好みである)の 4 段階で尋ねておく.

推薦者ありの実験では, 1 つの推薦リストにおいて, 4 種類の推薦者が 2 本ずつ映画を推薦する. 推薦者の種類ごとに推薦リストの順位におけるカウンターバランスをとる必要がある. そこで, 4 種類の推薦者が映画リストの 1～8 位に平等に配置されるよう,「1 位と 8 位」,「2 位と 7 位」,「3 位と 6 位」,「4 位と 5 位」の配置位置に 4 種類の推薦者を入れ替えて配置し, 合計 4 つの配置パターンで映画リストを作成する.

各被験者は 8 本の映画を含むリストで 4 回推薦を受けるため, 32 本の映画を準備する. 32 本の映画は, 上記で説明したアルゴリズムにより選択される. 評価値が 0(全く好みでない)と付与されたジャンルは, この 32 本の映画集合には含まれず, 評価値が 1～3 と付与されたジャンルは, 評価値が大きくなるほど多く含まれるように, 必要な映画の本数が決定される. その後, ジャンル分けされた映画群から重複なしで映画が必要な本数分ランダムに選択される.

グループ B の被験者の推薦リストを作成する手順について説明する. 32 本の映画を選択後, その映画を推薦リスト 1～4 のいずれかに割り当てる. 具体的には, まず 32 本の選択された映画は, ジャンル評価値が高かったジャンル順に整列される. なお, 同一ジャンル内では, ランダムに整列されている. この整列された映画群の上位から(すなわち 1 位～4 位)1 本ずつ推薦リスト 1～4 の順に割り当てる. 次に, 残りの映画群から(すなわち 5 位～8 位)1 本ずつ推薦リスト 4～1 の順に割り当てる. 残りの映画群に対しても同様に割り当てていく. これにより, 被験者が好きなジャンルの映画が特定のリストに偏らないようにする. リスト内の 8 本の映画は, 改めてランダムに並べられる.

グループ A の被験者の推薦リストを作成する手順について説明する. 推薦リストに割り当てる映画の選び方はグループ B の時と同様である. グループ A の被験者には, さらに推薦する各映画に対して推薦者を割り当てる. 4 つの推薦リストと前述した推薦者の配置パターンをランダムに組み合わせる.「映画コメンテーター」には 2 人を用意しているが, 推薦リストにおける映画コメンテーター用の配置位置(2 か所存在する)に, 2 人を 1 本ずつランダムに割り当てる.「友人」も, 推薦リストにおける友人用の配置位置(2 か所存在する)に, 異なる 2 名を割り当てる. この 2 名は, 該当する映画のジャンルを高く評価した同じ組織内の他の被験者からランダムに選択される.

2.4 実験方法

実験は, 互いにメンバーを認識している組織単位で行う. 具体的には, 10～20 名程度で構成される当大学商学部のいくつかのゼミを対象の組織とした. 具体的には, 3 年生の 4 つのゼミと 4 年生の 5 つのゼミに協力してもらい, 3 年生のゼミと 4 年生のゼミが均等になるように, ランダムにグループ A とグループ B に割り当てた. 詳細には, 3 年生の 2 つのゼミと 4 年生の 3 つのゼミをグループ A に, 3 年生の 2 つのゼミと 4 年生の 2 つのゼミをグループ B に割り当てた. 被験者は, 合計 134 名(男性 62 名, 女性 72 名)で, 内訳はグループ A が 74 名(男性 35 名, 女性 39 名), グループ B が 60 名(男性 27 名, 女性 33 名)である.

実験は以下の 3 つのステップで行われる.

(1) 事前調査(被験者による回答)
(2) 映画リストの作成(筆者らによる準備)
(3) 実験(被験者による回答)

(1)の事前調査では, 洋画を見る頻度を 5 段階(映画自体全く見ない, 洋画を全く見ない, あまり見ない, たまに見る, よく見る)で問う.「映画自体全く見ない」または「洋画を全く見ない」と答えた被験者は実験対象外となる. 次に, 14 種類の映画ジャンルに対して評価付けを 4 段階(興味なし, 少し好き, まあまあ好き, とても好き)で行ってもらう. また, 被験者には高度な推薦システムによる推薦であるように見せかけるために, 被験者が見たことのある映画 10 本を 5 段階の星により評価してもらう. また, 他のメンバーの推薦リストに推薦者として表示されるときのアバターの画像を設定してもらう(肌の色や髪型, メガネの有無など). 最後に, 被験者情報(ゼミ名, ログイン情報, 氏名)が入力され, 終了する. なお, 氏名は他のメンバーの推薦リストに推薦者として提示されるときに表示される.

(2)の映画リストの作成は, (1)で取得した情報を用いて, 前節で説明した方法で, 各被験者に映画リストを作成する. これはオフラインのプログラムにて行われる.

(3)の実験では, 被験者は実際に実験用推薦システムを利用する. また, 実験後にいくつかの調査用アンケートに回答する. まず, 被験者はシステムにログインし, 実験の説明を受ける. ここでは, 映画は 1 本 300 円で 48 時間見放題で, 1 ヶ月以内に見たい映画があれば購入するように指示する. 購入する映画の本数に制限はない. 推薦者ありのグループ A には, 推薦者の説明も提示する(図 1 参照). また, 推薦者ありのグループ A では, 推薦者を提示するインタフェース(図 2-(a)参照)にて推

(a) 映画ロボット　(b) 映画コメンテーター　(c) 友人

図 1 推薦者の紹介文

(a) 推薦者ありの場合　(b) 推薦者なしの場合　(c) カートの中身

図 2 推薦システムのインタフェース(スマートフォンの場合)

薦を受ける. 推薦者なしのグループ B では, 同じインタフェースで推薦を受けるが, 推薦者は表示されない(図 2-(b)参照). 提示される推薦リストは(2)で作成されたものである. 推薦リストは 4 回提示される. 各回で表示される映画リストの中に鑑賞したい映画があれば, 被験者はカートに追加し購入する(図 2-(c)参照). この行為を受容とみなす.

映画が推薦された後(推薦リストを確認し, 鑑賞したい映画を購入した後), 直前の推薦リストについて 3 つの項目を尋ねる. 1 つ目では, 推薦リストの有用度を 5 段階(全く有用でなかった, あまり有用でなかった, 少し有用であった, まあまあ有用であった, とても有用であった)で尋ねる. 2 つ目では, 各 8 本の映画が好みに合っていたかを 3 段階(合っていなかった, どちらとも言えない, 合っていた)で尋ねる. 3 つ目では, 被験者が見たことのある映画をチェックボックスにて選択してもらう.

4 回分の推薦を受けてもらった後, システム自体への信頼度を 5 段階(全く信頼しない, あまり信頼しない, 少し信頼する, まあまあ信頼する, とても信頼する)で答えてもらう. RS への信頼の程度は, 推薦結果の採用に関する意思決定プロセスにおいて, 重要な役割を果たすため[13], この質問を加えた.

また, グループ A には, 推薦者表示がどの程度参考になったかと, 各推薦者がどの程度参考になったかを 5 段階(全く参考にならなかった, あまり参考にならなかった, 少し参考になった, まあまあ参考になった, とても参考になった)で評価してもらった. また, その評価の理由も自由記述にて入力してもらった.

3 実験結果

3.1 受容率

分析には, 各被験者のアイテム受容率を計算して用いる. 受容率は以下の式で求められる.

受容率 = 受容された数 / 提示した数

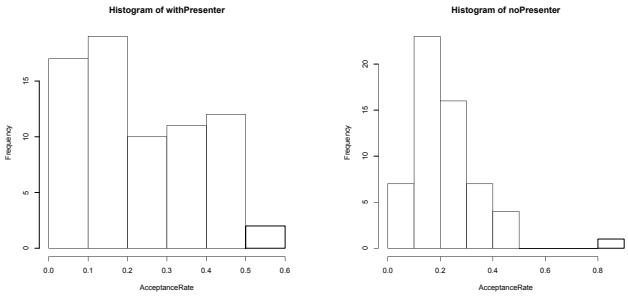
不誠実な実験参加者を除くため, それぞれの推薦リストにおける評価値の入力またはアンケートにおいて, 全項目のうち 8 割に同じ値を入力した被験者(具体的には, 好みに合うかの質問について, 32 本のうち 25 本以上に同じ回答をした被験者)を抽出し, その中から同じ値が連続している被験者(例えば, 1 回目の推薦リストでは値がばらついているが, 2 回目以降の推薦リストでは全部同じ値を入れているような被験者)を除外した. その結果, 分析対象となったのは, 推薦者ありで 71 名, 推

薦者なしで 58 名となった.

分析対象となったデータ(評価指標)について, シャピロ・ウィルク検定を行ったところ, 全て正規分布に従わなかったため(p-value < 0.05), ノンパラメトリック検定を採用することにした.

3.2 推薦者ありとなしの比較

推薦者ありの推薦リストと, 推薦者なしの推薦リストにおいて, 受容率に差が出るかを検証した. 推薦者ありと推薦者なしの受容率のヒストグラムを図 3 に示す. この図から, 推薦者ありでは分布がより広い範囲に及んでいることが分かる. 等分散性の検定を行ったところ, 有意差はなかった(p-value = 0.6287). マン・ホイットニーの U 検定によって, 代表値の差の検定も行ったが, 有意差はなかった(p-value = 0.7627).



(a) 推薦者あり (b) 推薦者あり

図 3 受容率のヒストグラム

3.3 推薦者の種類の比較

推薦者ありのデータ(71 名分)で, 推薦者の種類によって受容率に差が出るかどうかを検証した. ここでの受容率の計算は, 推薦者の種類ごとに計算した. すなわち注目する種類の推薦者が推薦した映画に対して計算した.

推薦者の種類と受容率に関係があるか否かを確かめるために, クラスカル・ウォリス検定を行った. その結果, 推薦者の種類において有意差が確かめられた(p-value = 0.0150*). 次に, 推薦者の種類間で受容率に差があるか否かを確かめるために, ホルム法によるマン・ホイットニーの U 検定で多重比較を行った. その結果,「友人と映画ロボット」と「友人と推薦者なし」で有意差がみられた(p-value = 0.0201*, p-value = 0.0042**).

さらに, 被験者の好みに合った映画で, なおかつ見たことがない映画に限定して分析を行った. クラスカル・ウォリス検定を行ったところ, 有意差が確かめられた(p-value = 0.0272*). また, ホルム法によるマン・ホイットニーの U 検定で多重比較を行ったところ,「友人と映画ロボット」と「友人と推薦者なし」の間で差があるとわかった(p-value = 0.0280*, p-value = 0.0280*).

また, 推薦者の種類と受容の関係を調べるために表 1 のデータでカイ二乗検定も行った. カイ二乗検定では, 有意差は見られなかったが(p-value = 0.0747), 残差分析の結果,「友人」のみ調整済み残差の絶対値が 2.5798 で 1.96 を超えたため, 期待度数とは大きく異なる観測度数であることがわかった. よって特に「友人」は受容率に影響を持つと考えられる.

表 1 クロス集計表

	F	E	R	N
受容された本数	111	79	77	71
受容されなかった本数	33	39	42	40

F:友人, E:映画コメンテーター

R:映画ロボット, N:推薦者なし

3.4 考察

推薦者ありの推薦リストと推薦者なしの推薦リストの受容率には, 統計的有意差はなかった. しかし, 推薦者ありの受容率の分布には, ややばらつきが見られ, また 3.3 節の推薦者の種類の比較では, 推薦者の種類によって受容率に違いが見られたことから, 複数の推薦者の種類の存在が, 全体の受容率のばらつきを引き起こしたと考えられる.

推薦者の種類の比較から, 特に「友人」は推薦者として大きな影響力を持つことがわかった. グループ A に実験終了時に尋ねた, 各推薦者がどの程度参考になったかを問う質問に対する自由記述での回答では,「友人」をより信頼したという肯定的な意見が多くあり, 反対に「映画コメンテーター」や「映画ロボット」は, 信頼できる存在なのか分からないという内容の意見があった. よって, より信頼がおける「友人」からの推薦は多くの被験者に受け入れられたと考えられる.

Gellerstedt と Arvemo の研究[11]では, オンラインの口コミにおいて, 実世界の友人の影響力の強さを明らかにしていたが, 推薦システムにおいても支持されることが, 本研究の結果で示された. 一方, Bonhard と Sasse の研究[9]や Zhou と Duan の研究[12]で明らかにされていた人間の専門家による意思決定への影響力は, 本研究では確かめられなかった. この理由としては, 人間の専門家よりも影響力の強い「友人」の推薦も組み入れたことで, 相対的に専門家の影響力が低下したものと思われる. また, レビューと推薦者提示という情報の違いも, 専門家の影響力が上がらなかった原因と考えられる.

また, 推薦システム自体に被験者がどの程度信頼しているかも確かめた. 実験の最後に設けた, 実験システム自体への信頼度を問う質問の回答値を, 推薦者ありと推薦者なしで比較する. マン・ホイットニーの U 検定による代表値の差の検定の結果, 有意差が確かめられた

(p-value = 0.0064**). 回答値の平均は推薦者ありが 3.126, 推薦者なしが 2.724 で, 推薦者ありの方がシステムに対する信頼度が高い. 各推薦者がどの程度参考になったかを問う質問に対する自由記述の回答では,「他人の意見は面白い」や「他人からの推薦は説得力があり興味が湧く」という意見があった. このことから, 推薦者ありでは, ユーザはより良い推薦の体験ができた可能性がある.

3.5 制約

実験の結果から,「友人」は受容率に影響を持つことがわかった.「映画コメンテーター」には, 映画に詳しい芸能人で, 日本で最も有名な「有村昆」と「LiLiCo」を採用した. これは多くの被験者が認知しており, 映画通であると認識していることを期待してのことである. しかし, その認識の程度は, 被験者によって異なった可能性は否めない. このような実在する映画評論家を使う方法以外にも, 実用性は劣るかもしれないが, 採用する映画評論家についての説明を行い, プライミングを行う方法も考えられる.

また, 本実験では, 組織として大学のゼミを扱った. そのため, 組織内のメンバー構造がフラットで, なおかつ同じ世代の人間になった. 一般の職場のように, 多くの年代のメンバーが所属する組織を対象に実験を行うと, 異なる結果になるかもしれない. 対象組織の違いに対する実験は, 今後の課題となる.

4 おわりに

本研究では, RS での推薦者の有無や種類がユーザのアイテム受容に与える影響を調査した. 被験者実験の結果, 推薦者ありと推薦者なしの推薦リストにおいて, アイテム受容に統計的な有意差はないが, 推薦者ありだと受容のされやすさに, 若干のばらつきが生じる傾向があった. また, 4 種類の推薦者の中で「友人」がより強い影響力を持ち, 実世界で積み重ねてきた信頼関係が重要な働きをする可能性があるとわかった. 現在, 用いられている多くの商用の推薦システムでは, 推薦するアイテムの情報を推薦リストに表示するものの, ある特定の推薦者を合わせて提示するものではない. ビッグデータに基づくコンピュータのアルゴリズムによる推薦と, 個別のユーザに注目した人物による推薦とを組み合わせることで, より良いユーザ体験を提供し, 推薦の受容率を高めることができる可能性がある. 今後は, 映画評論家の設定にプライミングを行い, 推薦者の設定をより平等にして実験を行いたい. また, 推薦者の有無や種類がユーザのアイテム受容に与える影響と, ユーザのパーソナリティの関係に注目した調査を行いたい.

謝辞

本研究は JSPS 科研費(19K12242)の助成を受けたものである.

参考文献

[1] Goldberg, D., et al.: Using Collaborative Filtering to Weave an Information Tapestry, Communications of the ACM, 35(12), pp. 61-70, 1992.

[2] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J. T.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews, in Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW'94), pp. 175-186, 1994.

[3] Shardanand, U. and Maes, P.: Social Information Filtering: Algorithm for Automating Word of Mouth, in Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'95), pp.210-217, 1995.

[4] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T.: Evaluating Collaborative Filtering Recommender Systems, ACM Transactions on Information Systems (TOIS), 22(1), pp. 5-53, 2014.

[5] Konstan, J. A. and Riedl, J. T.: Recommender Systems: from Algorithms to User Experience, User Modeling and User-Adapted Interaction, 22(1-2), pp. 101-123, 2012.

[6] Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H. and Newell, C.: Explaining the User Experience of Recommender Systems, User Modeling and User-Adapted Interaction, 22(4-5), pp. 441-504, 2012.

[7] Konstan, J. A and Riedl, J.: Recommender Systems: Collaborating in Commerce and Communities. In ACM CHI, 2003.

[8] Staffa, M. and Rossi, S.: Recommender Interfaces: The More Human-Like, the More Humans Like. In: Agah, A., Cabibihan, J. J., Howard, A., Salichs, M., He, H. (eds) Social Robotics. ICSR 2016. Lecture Notes in Computer Science, Vol. 9979. Springer, Cham, 2016.

[9] Bonhard, P. and Sasse, M. A.: "I thought it was terrible and everyone else loved it" — A New Perspective for Effective Recommender System Design. In: McEwan T., Gulliksen J., Benyon D. (eds) People and Computers XIX — The Bigger Picture. Springer, London, 2006.

[10] Dimensional Research. Survey: 90% of Customers Say Buying Decisions are Influenced by Online Reviews, 2013.

[11] Gellerstedt, M. and Arvemo, T.: The Impact of Word of Mouth When Booking a Hotel: Could a Good Friend's Opinion Outweigh the Online Majority? Inf Technol Tourism 21, pp. 289-311, 2019.

[12] Zhou, W. and Duan, W.: Do Professional Reviews Affect Online User Choices Through User Reviews? An Empirical Study, Journal of Management Information Systems Published Online, 2016.

[13] Jameson, A. et al.: Human Decision Making and Recommender Systems. In Recommender Systems Handbook, pp. 611-648, 2015.

ARG WI2 No.16, 2020
WI2-2020-12

アイテム分散表現の階層化・集約演算に基づくセッションベース推薦システム

棚木 佑真[†,a]　　岡本 一志[††,b]

†電気通信大学 情報理工学域　††電気通信大学 大学院情報理工学研究科 情報学専攻

a) *n1710462@edu.cc.uec.ac.jp*　b) *kazushi@uec.ac.jp*

概要　本研究では，セッションデータに対して Item2Vec によるアイテム分散表現を導入した推薦システムを提案する．さらに，セッションデータに特化したユーザ分散表現 (リアルタイムユーザ表現) を推薦時に導入することで，推薦の精度の向上を目指す．リアルタイムユーザ表現は推薦のタイミングに応じて，単純な計算でリアルタイムに構築されるという特徴を持っており，ユーザ嗜好のリアルタイムな追跡を試みている．推薦アイテムの探索時にはリアルタイムユーザ表現に加え，他のユーザ分散表現も考慮している．実験結果として，提案手法の Recall@k・MRR@k は Item2Vec よりも高く，FPMC の MRR@k と同等のスコアになることを確認している．また FPMC と比較して，提案手法はモデル訓練が短時間で可能なことを確認している．

キーワード　情報推薦，協調フィルタリング，セッションデータ，分散表現，近傍法

1　はじめに

Session-Based Recommender Systems(SBRS) とは，部分的に既知のセッション情報が与えられたとき，セッション内・セッション間に潜在する複雑な関係をモデル化し，未知のセッション情報を予測する推薦システムである [1]．アイテム分散表現を活用した推薦システムの研究が近年行われているが，SBRS にアイテム分散表現を取り入れた推薦システムの研究は限られている．ユーザの分散表現の導入により推薦精度の向上が期待されるが，ユーザの嗜好は変化が激しく，頻繁な分散表現の再構築が必要になり，計算コストが高くなる問題がある．

本研究では，セッションデータに対してアイテム分散表現を導入した推薦システムを提案する．アイテム分散表現の構築には自然言語処理分野における実績がある Word2Vec[2] を商品閲覧履歴に応用した Item2Vec[3] を利用する．さらに，セッションデータに特化したユーザ分散表現（リアルタイムユーザ表現）を推薦時に導入することで，推薦の精度の向上を目指す．リアルタイムユーザ表現は推薦のタイミングに応じて，単純な計算でリアルタイムに構築できる特徴があり，計算コストの問題解決を試みる．

評価実験では提案手法とその関連手法によって，ターゲットアイテムに対する推薦アイテム集合 R をそれぞれ生成する．R に対して，Recall@k, MRR(Mean Reciprocal Rank)@k といった評価指標によって推薦精度を評価することで，セッションデータに対する有効性を検証する．

Copyright is held by the author(s).
The article has been published without reviewing.

2　関連研究

Wang らは SBRS がモデル化すべき重要な要素として，

(a) ユーザの連続的な行動: 最新セッションにおいて出現したアイテムおよびその順序

(b) ユーザの一般的な嗜好: ユーザが典型的にどのようなアイテムを好むか

の 2 点を挙げている [4]．本研究はこの 2 つの要素の重要性を支持しており，このことを軸に既存手法の分析を進めていき，本研究で提案する手法の妥当性を検討する．

Word2Vec[2] を商品購買履歴に応用した Item2Vec[3] によるアイテム分散表現は，同一ジャンルのアイテム間のコサイン類似度が高くなるなど，アイテム間の関係を推測できる．セッションデータにおいても，連続するアイテム間には関係性があることが仮定でき，Item2Vec を導入できることが予想される．そこでまず，直前のアイテム分散表現に対する最近傍 k アイテムを推薦アイテムとする手法が考えられるが，アイテム同士の共起性しか考慮しておらず，(a) も (b) もモデル化できていない．

その一方で，Grbovic らの user2vec[5] は，ユーザとアイテムの分散表現を同時に学習する．ユーザ分散表現の学習には Paragraph2Vec[6] を利用し，この表現の導入により高い精度を確認している．このユーザ分散表現は (b) のモデル化の役割を果たしている．しかし，Grbovic らの分析において，ユーザの嗜好の変化が激しく，頻繁な再構築が必要であることが示唆されており [5]，計算コスト面での課題がある．本研究もユーザ分散表現を導入するが，Paragraph2Vec を利用するのではなく，アイテム分散表現を利用した単純な計算による構築により，計算コストの削減を目指す．

また一方で，Factorizing Personalized Markov Chains (FPMC)[7] モデルは推薦システムの一般的な手法である行列因子分解 (MF) とマルコフ連鎖 (MC) を組み合わせており，両者の利点を兼ね備えている．ここで，MF が (a)，MC が (b) の役割を果たしており，SBRS がモデル化すべき重要な要素のいずれも満たしている．しかし，計算コストが大きいことや，ユーザ情報が欠損していると推薦が困難になるといった欠点がある．

最後に，多層ニューラルネットワーク (NN) モデルを利用した研究も提案されている．代表的なモデルとして，GRU4Rec[8] がある．これらは (a) のモデル化に軸を置いており，高い精度を確認している．しかし，計算コストが大きい問題や出現順序を厳密に考慮し過ぎているといった指摘がなされている [9].

本研究で提案する手法は，セッションの分散表現（セッション表現）が (a)，ユーザの分散表現（ユーザ表現）が (b) のモデル化の役割を果たすようにし，2 つの重要な要素を満たす．また，浅い NN モデルの Item2Vec がベースであるため，低い計算コストで済むことが期待できる．

3 提案手法

アイテム集合を V，セッション集合を S，ユーザ集合を U とする．本研究で設定するタスクでは，ユーザ $u \in U$ に対して，閲覧したアイテムを時系列順で並べたセッションの集合 $S_u = \{s_u^1, s_u^2, \cdots, s_u^i\} \subset S$ $(s_u^t \in V \times V \times \cdots \times V)$ が与えられたとき，次に閲覧するアイテム v_t をターゲットとして予測する．この予測のために推薦システムはランク付けされた推薦アイテム集合 R を生成する．

提案手法は，分散表現の次元数を d とすると，

〈1〉 あらかじめ，過去一定期間のセッション集合 $S_p \subset S$ を使って，アイテム $v \in V$ の分散表現 $\boldsymbol{x}_v \in \mathbb{R}^d$ を Item2Vec により構築する．

〈2〉 セッション $s = s_u^t$ に対して，アイテムの閲覧ごとに分散表現 $\boldsymbol{x}_s \in \mathbb{R}^d$ を構築/更新する．

〈3〉 セッション $s = s_u^t$ の終了時に，ユーザ u の分散表現 $\boldsymbol{x}_u \in \mathbb{R}^d$ を構築/更新する．

〈4〉 ユーザ u に対する R の生成が必要となった時点で，$\boldsymbol{x}_{s_l}$ と $\boldsymbol{x}_u$ を使い，リアルタイムユーザ表現 $\boldsymbol{z}_u \in \mathbb{R}^d$ を構築する．なお，$s_l = s_u^i \in S_u$ である．

〈5〉 $\boldsymbol{z}_u$ を使って各分散表現の k 近傍探索を行い，R を生成する．詳細は後述する．

の推薦プロセスになっている．

〈2〉 について，$\boldsymbol{x}_s$ $(s = s_u^t)$ は $\boldsymbol{x}_v$ と $\boldsymbol{x}'_s$（更新前の $\boldsymbol{x}_s$）のコサイン類似度 (COS) または順序差減衰 (ODD: Order Decay Difference) によって重み w_s を計算し，w_s を使った加重平均・加重和により構築する．w_s の計算方法は

$$w_s = \left|\cos(\boldsymbol{x}'_s, \boldsymbol{x}_u)\right| \quad (1)$$

$$w_s = \exp(-\lambda) \quad (2)$$

の 2 通りである．なお，λ は崩壊定数というパラメータである．$\cos(\boldsymbol{x}_1, \boldsymbol{x}_2)$ は $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ のコサイン類似度である．本研究では，式 (1),(2) を使ったモデルを，proposal-COS/ODD とする．w_s を使って $\boldsymbol{x}_s$ を更新する．ここでは

$$\boldsymbol{x}_s = w_s \boldsymbol{x}'_s + (1 - w_s)\boldsymbol{x}_v \quad (3)$$

$$\boldsymbol{x}_s = (1 - w_s)\boldsymbol{x}'_s + w_s \boldsymbol{x}_v \quad (4)$$

$$\boldsymbol{x}_s = w_s \boldsymbol{x}'_s + \boldsymbol{x}_v \quad (5)$$

$$\boldsymbol{x}_s = \boldsymbol{x}'_s + w_s \boldsymbol{x}_v \quad (6)$$

の 4 通りの計算方法を考える．ただし，式 (2) の場合は式 (5) で固定する．式 (3),(4) はそれぞれ，セッション/アイテムを軸と（係数を重み w_s に設定）した加重平均，式 (5),(6) は加重和を表している．

〈3〉 について，$\boldsymbol{x}_u$ は $\boldsymbol{x}_{s_l}$ と $\boldsymbol{x}_{s \in S_u}$ のコサイン類似度の絶対値 $w_{s \in S_u}$ に基づく $s \in S_u$ の加重平均・加重和により構築する．$w_{s \in S_u}$ の計算方法は

$$w_{s \in S_u} = \left|\cos(\boldsymbol{x}_{s_l}, \boldsymbol{x}_{s \in S_u})\right| \quad (7)$$

の通りである．$w_{s \in S_u}$ を使って $\boldsymbol{x}_u$ を更新する．ここでは

$$\boldsymbol{x}_u = \frac{\sum_{s \in S_u} (w_s \boldsymbol{x}_s)}{\sum_{s \in S_u} w_s} \quad (8)$$

$$\boldsymbol{x}_u = \sum_{s \in S_u} (w_s \boldsymbol{x}_s) \quad (9)$$

の 2 通りの計算方法を考える．

〈4〉 について，$\boldsymbol{z}_u$ は $\boldsymbol{x}_{s_l}$ と $\boldsymbol{x}_u$ のコサイン類似度の絶対値 w_z に基づく加重平均・加重和により構築する．w_z の計算方法は

$$w_z = \max\left\{\left|\cos(\boldsymbol{x}_{s_l}, \boldsymbol{x}_u)\right|, b\right\} \quad (10)$$

の通りである．なお，b は最低推薦貢献度と呼び，セッション表現・ユーザ表現のいずれかに大きく依存することを避けるための閾値である．

w_z を使って $\boldsymbol{z}_u$ を構築する．ここでは

$$\boldsymbol{z}_u = w_z \boldsymbol{x}_{s_l} + (1 - w_z)\boldsymbol{x}_u \quad (11)$$

$$\boldsymbol{z}_u = (1 - w_z)\boldsymbol{x}_{s_l} + w_z \boldsymbol{x}_u \quad (12)$$

$$\boldsymbol{z}_u = w_z \boldsymbol{x}_{s_l} + \boldsymbol{x}_u \quad (13)$$

$$\boldsymbol{z}_u = \boldsymbol{x}_{s_l} + w_z \boldsymbol{x}_u \quad (14)$$

の 4 通りの計算方法を考える．式 (11),(12) はそれぞれ，セッション/ユーザを軸と（係数を重み w_z に設定）した加重平均であり，式 (13),(14) は加重和を表している．

ここで，2 節で挙げた SBRS がモデル化すべき重要な 2 つの要素と関連付けると，$\boldsymbol{x}_{s_l}$ は (a), $\boldsymbol{x}_u$ は (b) のモデル化を試みており，$\boldsymbol{z}_u$ は (a) と (b) の両方をモデル化を試みたハイブリッドな表現である．

各分散表現は Item2Vec によるアイテム分散表現がベースであり，これらは同じベクトル空間を共有している．よって，各分散表現に対して異なるタイプの分散表現の k 近傍探索ができる．提案手法はこの利点を活用して，

〈5-1〉$\boldsymbol{z}_u$ に対して，近傍アイテム集合 $R_1 \subset V$ ($|R_1| = k_v$) と近傍ユーザ集合 $U_n \subset U$ ($|U_n| = k_u$) を作成する．

〈5-2〉$\boldsymbol{x}_{u \in U_n}$ に対して，近傍アイテム集合 $V_u \subset V$ ($|V_u| = k_v$) を作り，$R_2 = \bigcup_{u \in U_n} V_u$ とする．

〈5-3〉$v \in R_1 \cap R_2$ を $\boldsymbol{z}_u$ に距離が近い順で R に追加する．

〈5-4〉$|R_1 \cap R_2| < k$ の場合，$v \in R_1 \setminus R_2$ を $\boldsymbol{z}_u$ に距離が近い順で $|R| = k$ になるまで R に追加する．

の手順により R を生成する．なお，距離関数としてユークリッド距離を適用する．

このように，提案手法はアイテム・セッション・ユーザと階層化するため，アイテム分散表現を使って様々な集約演算を行っていることが特徴である．

4 セッションデータを用いた評価実験

4.1 評価方法

実際のセッションデータにおいて，各手法によってそれぞれ v_t に対する R を生成し，その差異を評価することで，提案手法のセッションデータに対する有効性を確認する．本研究では，データセットとして Trivago[1] と DIGINETICA[2] を使う．Trivago はホテル検索サイト，DIGINETICA は EC サイトのアイテム閲覧履歴データである．前処理として，出現回数が 5 回未満のアイテムと，サイズ 1 のセッションを除外し，セッション内で同じアイテムが連続して出現する場合は 1 つにまとめる．処理後の各データセットの統計情報を表 1 に示す．

ベースラインとして，人気上位 k 個を R とする POP と，関連研究で挙げた直前アイテム分散表現の最近傍 k アイテムを R とする Item2Vec，FPMC を使う．

v_t はテストセット中の各セッションの終了直前のログのアイテムに設定する．各手法の評価には Recall@k と MRR@k を利用する．本実験では $k = 20$ とする．

4.2 評価結果

表 2 に各手法による推薦の評価結果を示す．まず Trivago の結果を見る．Recall と MRR ともに，提案手法は Item2Vec より高いスコアであり，Item2Vec によるアイテム分散表現を効果的に活用できると考える．proposal-COS の MRR は ODD よりも高く，ユーザの閲覧アイテムを高い自信で予測することが確認できる．Recall は FPMC が最も優れているが，MRR は proposal-COS が FPMC より僅かに上回っている．よって，FPMC は v_t のヒット率が高いが，ランキングで下位に出現しやすいとわかる．

次に DIGINETICA の結果を見る．Trivago と同様に提案手法が Item2Vec より優れている．しかし，proposal-ODD が COS より MRR が高く，Trivago とは異なる傾向である．また FPMC の MRR が Item2Vec よりも低いが，DIGINETICA にはユーザ ID が欠損したセッションが存在し，ユーザ情報を必須とする FPMC が適用できないからだと考える．一方，提案手法はユーザ情報が不足している場合，FPMC より MRR が減少せず，このようなセッションデータにも有効的な手法であると考える．

別の結果として，k を 1 から 20 まで増やした際のスコア推移を示す．ここでは Trivago における MRR の推移のみ，図 1 に示す．$k = 20$ のみを示した表 2 では，proposal-COS の MRR は FPMC より僅かに高い程度であったが，図 1 より，$k = 1$〜19 でも常に proposal-COS が FPMC より高く，その差は k が小さいほど広がっていることが確認できる．よって，proposal-COS は R のサイズが小さい場合でも有効的な手法である可能性がある．一方で proposal-ODD は常に FPMC よりスコアが低く，Trivago には不向きな手法ではないかと考える．

提案手法のスコアは MRR を基準に検証セットで最も優れたスコアとなった計算方法をテストセットに適用した際のスコアである．その計算方法を表 3 に示す．まず $\boldsymbol{x}_s$ について，いずれも w_s を $\boldsymbol{x}'_s$ の係数として与えるが，Trivago は加重平均，DIGINETICA は加重和による構築方法が最も優れており，異なる傾向である．次に $\boldsymbol{x}_u$ について，DIGINETICA の proposal-COS のみ加重和，それ以外は加重平均が最適であり，一般的な嗜好のモデル化を試みる場合は概して加重平均による構築が適して

表 1: データセット統計情報

データセット	アイテム	セッション	ユーザ
Trivago	100,750	307,518	268,912
DIGINETICA	41,123	192,464	53,913*

* 匿名ユーザは数えていない

表 2: 推薦精度 ($k = 20$) [%]

データセット	Trivago		DIGINETICA	
評価指標	Recall@k	MRR@k	Recall@k	MRR@k
POP	0.808	0.185	1.01	0.217
Item2Vec	27.7	7.30	26.4	7.52
FPMC	**43.5**	12.26	19.0	5.77
proposal-COS	31.7	**12.29**	29.1	7.72
proposal-ODD	31.8	8.92	**30.5**	**8.16**

[1] https://recsys.trivago.cloud/challenge/dataset/
[2] https://competitions.codalab.org/competitions/11161

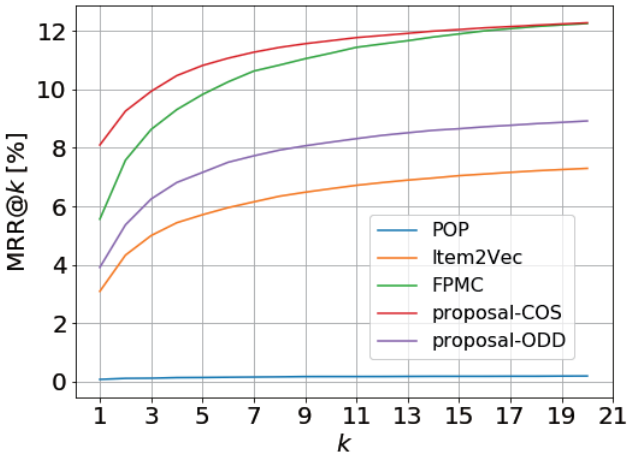


図 1: MRR@k による推薦評価 (Trivago)

表 3: 提案手法の最適な計算方法 (式番号)

データセット	モデル	x_s	x_u	z_u
Trivago	proposal-COS	(3)	(8)	(14)
Trivago	proposal-ODD	(5)*	(8)	(14)
DIGINETICA	proposal-COS	(5)	(9)	(11)
DIGINETICA	proposal-ODD	(5)*	(8)	(14)

* proposal-ODD は (5) で固定

いると考える．そして z_u については，DIGINETICA の proposal-COS のみセッションを軸とした加重平均，それ以外はユーザを軸とした加重和が良く，異なる傾向が見られる．データセットとモデルの組み合わせによって最適な構築方法が異なる要因は現状判明しておらず，今後考察を深めるべき点であると考える．

4.3 実行時間

各モデルの学習 (fit) および 1 ターゲットの予測 (predict) に要した時間を表 4 に示す．なお，POP を除き，各モデルはエポック数を 30 で揃えている．図 4 より，FPMC は他の手法と比較して，モデル学習に長時間要しており，実際に運用するにあたってサーバコストを大きくする必要がある．しかし，Item2Vec や提案手法といった Word2Vec ベースの手法は学習時間が FPMC と比較して極めて短く，モデルの頻繁な再学習が容易である．ただし，予測フェーズは，提案手法は FPMC より 1.5 倍前後の時間がかかる．これは予測時に k 近傍法によって都度アイテムやユーザの探索を行っているためであり，このアルゴリズム改善は今後の課題である．

5 おわりに

本研究では，セッションデータに対してアイテム分散表現を導入し，その階層化・集約演算によってリアルタイムユーザ表現を構築・推薦時に導入する SBRS を提案している．実験の結果，提案手法は Recall と MRR ともに Item2Vec より高く，FPMC と同等の MRR を獲得できることを確認している．今後は，各分散表現の構築・更新や推薦アイテム探索のアルゴリズム改善，多層 NN ベースのモデルとの比較，推薦アイテム集合の多様性や新規性を測ることができる指標を用いた各モデルの評価を行う予定である．

表 4: 各モデルの実行時間

データセット	Trivago		DIGINETICA	
フェーズ	fit [s]	predict [ms]	fit [s]	predict [ms]
POP	0.478	0.00886	0.177	0.00593
Item2Vec	133	8.44	56.1	6.18
FPMC	23647	73.8	2081	28.9
proposal-COS	176	113	76.6	90.1
proposal-ODD	137	135	70.9	76.0

謝辞

本研究は JSPS 科研費 JP18K18159 の助成を受けたものです．

参考文献

[1] S. Wang, C. Longbing, W. Yan: A Survey on Session-based Recommender Systems, arXiv, arXiv:1902.04864, 2019.

[2] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean: Distributed Representations of Words and Phrases and their Compositionality, *Proc. of Advances in Neural Information Processing Systems 26*, 3111—3119, 2013.

[3] O. Barkan, N. Koenigstein: Item2Vec, Neural Item Embedding for Collaborative Filtering, arXiv, arXiv:1603.04259, 2016.

[4] P. Wang, J. Guo, Y. Lan, J. Xu, S. Wan, X. Cheng: Learning Hierarchical Representation Model for Next Basket Recommendation, *Proc. of the 38th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 403–412, 2015.

[5] M. Grbovic, V. Radosavljevic, N. Djuric, N. Bhamidipati, J. Savla, V. Bhagwan, D. Sharp: E-commerce in your inbox: Product Recommendations at Scale, *Proc. of the 21st ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 1809–1818, 2015.

[6] Q. Le, T. Mikolov: Distributed Representations of Sentences and Documents, *Proc. of the 31st Int. Conf. on Machine Learning*, 1188–1196, 2014.

[7] S. Rendle, C. Freudenthaler, L. Schmidt-Thieme: Factorizing Personalized Markov Chains for Next-basket Recommendation, *Proc. of the 19th Int. Conf. on World Wide Web*, 811–820, 2010.

[8] B. Hidasi, A. Karatzoglou, L. Baltrunas, D. Tikk: Session-based Recommendations with Recurrent Neural Networks, arXiv, arXiv:1511.06939, 2015.

[9] L. Hu, L. Cao, S. Wang, G. Xu, J. Cao, Z. Gu: Diversifying Personalized Recommendation with User-session Context, *Proc. of the 26th Int. Joint Conf. on Artificial Intelligence*, 1858–1864, 2017.

ARG WI2 No. 16, 2020
WI2-2020-13

決定係数を用いたアイテム推薦理由の説明文の生成手法の検証

佐藤匠† 當間愛晃‡

†琉球大学大学院理工学研究科情報工学専攻 〒903-0213 沖縄県中頭郡西原町字千原 1 番地

‡琉球大学工学部知能情報コース 〒903-0213 沖縄県中頭郡西原町字千原 1 番地

Email: †k198578@ie.u-ryukyu.ac.jp ‡tnal@ie.u-ryukyu.ac.jp

概要 近年提案されている機械学習を用いた推薦モデルは解釈が難しいブラックボックスとなっているものが多く，学習済みモデルの状態から推薦理由を解釈することは困難である．先行研究として機械学習モデルの予測結果を解釈するアルゴリズムである LIME を推薦システムに適用し，解釈結果から推薦理由の説明をユーザに提示する手法が提案されている．LIME は近似により単純な線形回帰を作成することで重要な特徴量を推論する．この際に単純な線形回帰を使うために起きうる望ましくない近似の結果により誤った重要度を得る場合はないだろうか．LIME の推薦モデルへの適用による特徴量の重要度の算出を行い，LIME の線形回帰の決定係数を用いて，推薦結果の考察をする．

キーワード 情報推薦，機械学習

1 はじめに

昨今では機械学習の技術発展に伴い社会実装の期待が高まっている．世界的潮流に伴って総務省は「国際的な議論のための AI 開発ガイドライン案」[1]で「透明性の原則-開発者は，AI システムの入出力の検証可能性及び判断結果の説明可能性に留意する．」としている．

近年は行列分解や機械学習アルゴリズムを用いた新たな推薦モデルが多く提案されている．これらの最先端のアルゴリズムを使用した推薦システムは，より高い推薦精度を発揮するが，解釈性が低く，ユーザに対して説明を与えることが難しい．そこで行列分解や機械学習アルゴリズムを用いながら解釈性の高い推薦モデルを実現するための研究が行われている．

森澤ら[2]は，機械学習モデルの解釈を行うための手法として M．T．Ribeiro ら[3]によって提案された LIME (Local Interpretable Model-agnostic Explanations)を 推薦システムに組み込むことによって説明を生成する手法を提案した．LIME は，任意の学習済み機械学習モデルに対して，特定の推論結果における各特徴量の重要度を算出するアルゴリズムである．

LIME はランダムサンプリングをした後，線形回帰を行うことにより単純な線形回帰モデルを作成することで各特徴量の重要度を算出する． 一般的に線形回帰はモデルの単純さから、望ましくない近似になる場合がある．LIME は特定のデータやシステムに限定的な既存の手法と異なり，任意の学習済み機械学習モデルに応用可能であるとされるが，推薦という問題領域に対して近似的手法を用いることによる弊害はありえないのだろうか．

本研究では、LIME アルゴリズムを推薦モデルへ適応し，説明を付与する際の決定係数を用いて検証，考察する．

2 関連研究

本節では，近年の行列分解や機械学習を使用した推薦モデルに対する，ユーザへの説明可能性を高めるための研究について，モデルのアルゴリズム別に分類して説明する．

2.1 LIME を使用した推薦理由提示手法

本節では，森澤らが提案した，LIME を使用した推薦理由の提示手法について述べる．

2.2 LIME アルゴリズム

LIME は，任意の学習済み機械学習モデルに対して，推論の結果を説明するためのアルゴリズムとして，M．T．Ribeiro らによって提案された．LIME は，推論結果の説明として，推論結果に影響を与えた任意の個数の特徴量のラベルと重要度を出力する．特徴量の選定と重要度の算出には，学習済み機械学習モデルにおける特徴空間内の特徴ベクトルと出力値を，ランダムサンプリングによる解釈性の高い線形回

Copyright is held by the author(s).
The article has been published without reviewing.

帰モデルを用いて学習することによって行う．[3]では，分類問題を推論する機械学習モデルに対して，LIME を適用して説明を生成するための手法が提案されている．回帰モデルにも適用可能な LIME による説明生成のアルゴリズムを，アルゴリズム 1 に示す．

アルゴリズム 1　LIME を用いた説明の生成
入力: 説明対象のモデル f,
入力ベクトル x
入力: サンプル数 N,
類似度カーネル関数 π_x
入力: 説明に用いる特徴量の数 K
出力: 各特徴量の重要度 w

```
Z←{}
for i ∈ {1, 2, 3, . . . , N} do
    z' ← sample_around(x)
    Z ← Z ∪ ⟨z_i, f(z_i), π_x(z_i)⟩
end for
w←K-Lasso(Z, K)
return w
```

2.3 Lasso 回帰

LIME は線形回帰モデルを作成する際に K-Lasso 回帰を用いる．LIME の実装においては Lasso 回帰や Ridge 回帰と近い．Lasso 回帰は線形回帰の一つである．線形回帰はロジスティック回帰，決定木などと並んで説明可能性の高いモデルとされる．

2.4 森澤らの推薦理由提示手法

森澤らの行った推薦理由提示手法を図 1 の模式図に示す．森澤らは LIME を推薦モデルへ適応し，出力させた特徴量の重要度から最も重要であった特徴量を要ししておいた定型文に挿入し，「この映画はジャンル(Drama)であるためあなたにおすすめします」という形の説明文を生成し，推薦モデルユーザへ提供しアンケート評価による評価を行った．

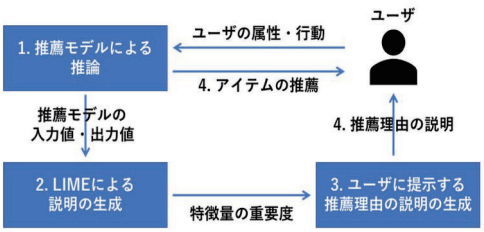


図 1: 森澤らの提案手法の模式図

3 実験手法

本稿では，森澤らの提唱した LIME を使用した推薦理由の提示手法に代表される，推薦という課題に対して単純な線形回帰を適応する際に起きうる望ましくない近似結果とはどのようなものかを，推薦モデルへ LIME を適応し，重要な特徴量を出力させた上で，LIME の決定係数などを踏まえて考察した結果について述べる．

3.1 使用する推薦アルゴリズム

推薦アルゴリズムは S．Rendle[5]によって提案された Factorization Machines (FM)を使用する．

入力ベクトル $x=(x_n+,...x_n,)\in R^n$ ，に対する評価推定値 $\hat{y}$ を，学習パラメータであるバイアス項 $w_0\in R$ ， $w=(w+,...w,)\in R$ と，変数間の相互作用項 $v_i=(v_{i1},+,...v_{i,k})\in R^k$ によって表現したモデルである．式(1)に示す．

$$\hat{y}(x)=w_0+\sum_{i=1}^{n} w_i x_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\langle v_i, v_j\rangle x_i x_j \quad (1)$$

3.2 使用するデータセット

データセットとして MovieLens 100k Dataset を用いた．946 人のユーザが 1682 本の視聴した映画に与えたスコアのデータセットで，ユーザ情報とアイテム情報，ユーザがつけた5段階評価の rating がおよそ 100，000 件含まれる．また，ユーザの年齢，性別，職業，視聴した年度，映画のジャンル情報などが含まれる．

3.3 特徴ベクトルの生成

推薦モデルへ入力する特徴量は 3.2 節のデータセットの行列を one-hot エンコードしたものを使用する．表1のように特徴量としてユーザ ID，映画 ID，リリース年度，年齢，性別，職業，ジャンル(18 種類)を one-hot 表現に変換した．

表 1: 使用した特徴量

名前	表現方法	次元数
ユーザ ID	one-hot encode	943
映画 ID	one-hot encode	1682
リリース年度	one-hot encode	71
年齢	0 から 1 に Scale	1
職業	one-hot encode	21
性別	女性 F,男性 M	2

ジャンル	one-hot encode	19

3.4 決定係数を用いた LIME の近似の精度の評価

LIME により学習済み推薦モデルの説明文を生成する際に LIME が生成する線形回帰モデルの決定係数を評価し，どのような場合において近似が望まない出力結果を得るのかを調べる．

実測値 y ，予測値 $\hat{y}$ ，実測値の平均 $\bar{y}$ としたとき，決定係数は式(2)のように表される．

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (2)$$

これを用いて，LIME が近似により説明としては不適切な出力結果を出力したかどうかを判断し，その具体的な評価をする．

4 実験結果

4.1 ヒストグラム

LIME にテストデータ 10000 件を入力し，特徴量の重要度を推論させ，その際の決定係数をヒストグラムたものを図2に示す．

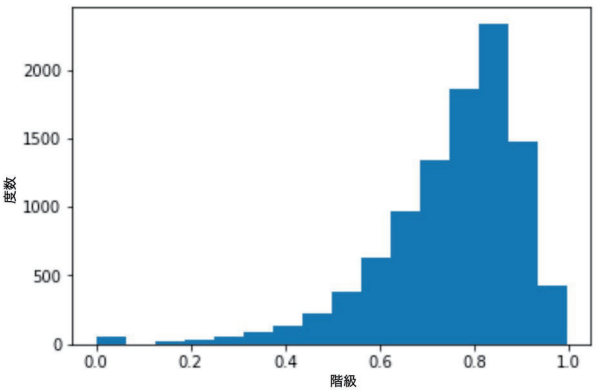


図 2: LIME の決定係数のヒストグラム

4.2 度数分布表

同じく，決定係数を度数分布表で示す．

表 2: 決定係数の度数分布表

	start	end	count
0	0	0.06	46
1	0.06	0.12	1
2	0.12	0.19	19
3	0.19	0.25	32
4	0.25	0.31	45
5	0.31	0.37	90
6	0.37	0.44	137
7	0.44	0.5	212
8	0.5	0.56	363
9	0.56	0.62	643
10	0.62	0.68	950
11	0.68	0.75	1346
12	0.75	0.81	1850
13	0.81	0.87	2370
14	0.87	0.93	1478
15	0.93	1	415

表 3: 決定係数の分布の詳細

	point
データ数	10000
平均	0. 753203
標準偏差	0. 147565
最小値	0. 000045
25%	0. 682564
50%	0. 786878
75%	0. 854717
最大値	0. 996186

4.3 精度の評価

決定係数の評価については一定の境があるわけではないが，一般的に決定係数が低く，0.6 以下などであれば学習が不適切であるという疑いがあり，1.0 に近い場合，過学習を疑いがあるといえる．

今回の実験結果から以下のことが言える．近似により一定数，LIME が適切に近似できていないため決定係数が低くなる場合が一定数みられる．しかしヒストグラムは右寄りの図形を示しており，全体の傾向としては高い値が得られている．平均は 0.75 と高く最小値は 0.000045 と極端に低い．表 2 より 0.00 から 0.06 の値のサンプル数は 46 でありその周辺の値より数が多い．0.12 からは徐々に数が増える分布になっている．表2から四分位数でみれば全体の 25%時点ですでに 0.68 近い数値が出ている．

4.4 個別例

決定係数に値が低いテストユーザの例を見て，どの

ような学習結果が LIME から得られているかを確認する.

決定係数の低かったユーザ ID 481 番に対するアイテム ID 500 番の推薦結果を該当ユーザの事例としてアイテム推薦した結果について詳しく述べる. 該当ユーザは決定係数は約 0.073 であった. 決定係数 0.07 は表 2 を確認するとこの該当ユーザ一例のみである. 推薦モデルは該当ユーザに対し約 3.939 という推薦結果を提示し, 実際の正解ラベルは 4 であった. 該当ユーザは 56 件のレコードがあり, 学習データの不足はないと思われる. ジャンル Adventure を含む映画を該当ユーザが視聴した履歴は全部で 7 つあり, そのうち Children's を含むものは該当の事例のみであった. 7 件とも 4 以上と高い評価をつけており, 2件は5である. 該当ユーザが Children's を評価した履歴は 3 つあり, 3 点が 1 つ 4 点が 2 つであった. アイテム ID 500 番は ”Fly Away Home “(1996) であり, Adventure,Children's のジャンル情報を持っていた. 全映画のうち Adventure,Children's を持つ映画は 42 件あり, 数としては少なくない.

よってこれらの結果から, 特徴量 Adventure は結果を高く評価する方向へ作用していると推論するのが妥当である.

表 4 に特徴量の重要度は出力されており, これを確認する. 正の重みを持つ特徴は, 出力値を正の値にするために寄与した特徴であり, 負の重みを持つ特徴は, 出力値を負の値にするために寄与した特徴であると解釈される. この場合, 特に Adventure は負の方向へ, Children`s が正の方向へ作用していると解釈できる. これは先述した該当ユーザの履歴の理解とは反する. また, 決定係数も低いため決定係数に従えば結果は説明として参考にするのには不適切であるといえる.

次に推薦理由の説明として評価する.森澤らの推薦理由の説明をする手順に従えば, 表 4 の重要度のスコアから高い重要度のスコアを用いてそれぞれの特徴量の推薦理由に対し用意しておいた定型文に挿入する. 該当ユーザの事例であればもっとも高い値はジャンル情報であるため「この映画はジャンル"Children's”であるため, あなたにおすすめします」となる. Children's を該当ユーザは 3 件視聴しており, 高いスコアを評価しているわけではないため, 推薦理由としては評価できない.

表 4: 該当ユーザの特徴量の重要度

特徴量のタグ	重要度のスコア
release_year=1996	-0.11156219773630258
year=1998	-0.06671403591467708
sex=M	-0.05242530014669192
Children's	0.023566247090890922
occupation=retired	0.012558840356626296
movie_id=500	0.003105300470527567
user_id=481	0.0030222771556819073
age	0.00012698185241982903
Animation	0.0

5 おわりに

本稿では, 推薦モデルに対して推薦理由を生成するために LIME を用いた際に単純な線形回帰の近似による望まない出力結果が得られる場合について決定係数を用いて考察を行った. ヒストグラム, 度数分布表などで決定係数の分布がどの様になっているかを可視化した. 個別具体的な例を取り上げ, どのような事例で望まない出力結果が得られるかを確認した. 決定係数が低いサンプル 1 つを確認する限りでは適切な結果であった.

今後の予定としては, 決定係数ではなく AIC(赤池情報量規準)といった他の精度評価の手法を用いて LIME がモデルへ当てはまっているのかを検証する実験を行うこと, また LIME ではない特徴量選択の手法を用いた場合について検証する実験などを検討している.

参考文献

1 総務省,国際的な議論のための AI ネットワーク社会推進会議平成 29 年 7 月 28 日,https://www.soumu.go.jp/main_content/000499625.pdf, 2020 年 11 月 10 日

2 森澤 竣, 山名 早人:機械学習モデルの解釈手法を用いたアイテム推薦理由の説明文の生成, DEIM2020 E6-2

3 M. T. Ribeiro, S. Singh, and C. Guestrin, :Why Should I Trust You?' Explaining the Predictions of Any Classifier, , in Proc. of ACM KDD, pp. 1135– 1144, 2016.

4 B. Abdollahi and O. Nasraoui, :Using Explainability for Constrained Matrix Factorization, in Proc. of ACM RecSys, pp. 79–83, 2017.

5 S. Rendle, “Factorization Machines, ” in Proc. of IEEE ICDM, pp. 995–1000, 2010.

ARG WI2 No.16, 2020
WI2-2020-14

深層学習による少数学習データでの2次元データの高品質化手法の提案

石原 正敏[†,a]　荒木 徹也[††,b]　石川 博[†,c]

† 東京都立大学大学院システムデザイン学部情報科学域　†† 群馬大学理工学部電子情報理工学科

a) *ishihara-masatoshi@ed.tmu.ac.jp*　b) *tetsuya.araki@gunma-u.ac.jp*　c) *ishikawa-hiroshi@tmu.ac.jp*

概要　本稿では超解像とガウシアンノイズ除去の二つの観点からデータ高品質化を行う．近年, 火星地表画像の超解像やヒートマップの高品質化など, 様々なデータの高品質化が求められる．提案手法の汎用性の評価の為,DIV2K 画像データセットを利用する．画像データセットから,128 × 128 × 3 のカラーの部分画像をランダムに切り出したものを学習に使用する．過剰適合を防ぎつつ精度を向上させる為, 入出力に直通の迂回路を設置することで過剰適応と精度の両立を行う．解像度の向上とノイズ除去では適切な手法が異なる為, 其々の手法を組み合わせることで両立する．この様な工夫により, 様々なケースに柔軟に対応した 2 次元データ高品質化が可能となる．

キーワード　深層学習, 超解像処理, ノイズ除去

1　はじめに

近年において観測したデータを有効活用するための課題として、二次元データの高品質化がある．取り扱うデータや観測方法によって, 特定のノイズが乗りやすさやデータの解像度が不十分であるなど, 適切な高品質化の方法は異なる．このように様々なケースに柔軟に対応するために, 機械学習によって高品質化を行う事が望ましい．

例えば, 高品質化のアプローチ一つに，単一画像超解像 (Single image super resolution) と呼ばれる解像度の低いデータを高解像化する手法がある．月面 DEM の高解像度化 [1] や深海海底地形図の作成 [2]，CT や MRI 画像の超解像 [3] が例として挙げられる．高品質化のための他のアプローチとして，ノイズ除去がある．ラマン散乱顕微鏡画像のノイズ除去 [4] や手書き文字画像のノイズ除去 [5] が例として挙げられる．複数の高品質化の為のアプローチに対応した機械学習手法を適応することで, 様々なケースに対応できる．

一方で常に十分な学習データが得られるとは限らない．例えば, 月面や海底の高度座標など, 限られた情報から高品質化の為の学習を行わなければならない場合がある．他にも画像の高品質化の場合は火星の地表や人体の患部など, どの学習データを利用したかが重要視される場合はビッグデータの転移学習などが好ましくない場合も存在する．学習データが不十分な場合, 過学習 (Overfitting) と呼ばれる訓練データに対して過剰に適合することで未知のデータに対応不能になる現象が発生する．一般に過学習対策手法は出力データの品質を劣化させるために, 品質劣化を抑える必要がある．

本稿では, 画像の超解像とガウシアンノイズ除去を行うことで, 二次元データの高品質化を行う少数データでも過学習を起しにくい機械学習手法を提案する．

2　関連研究

機械学習のアプローチの一つに，罰則と報酬によって神経接続 (Neural Net) を効率化させていくものがある．この方法は，汎用的な機械学習を実現する手段として注目されてきた [6]．畳み込みニューラルネットワーク (Convolutional Neural Network；CNN)(以降 CNN) と呼ばれる構造の登場は，画像処理に大きな影響を及ぼした [7]．

Nitish S., et al.[8] はランダムで神経細胞 (units) を不活化することで過学習抑制に高い効果を示す Dropout と呼称される手法を提案した．Dong ,C., et al.[9] は Super-Resolution Convolutional Neural Network(SRCNN) と呼ばれる 3 層か 4 層の CNN から構成される機械学習モデルを提案して, 高度な単一画像超解像を行った．当手法は低解像度画像を Bicubic 法で拡大したあとに当モデルを通すことで, 超解像を行った．Ronneberger, O., et al.[10] は U-Net と呼ばれる U 字型の構造を持つ機械学習モデルにより, 画像のセグメンテーションを行った．U-Net の U 字型の構造は, 全体的特徴と局所的特徴の両方を処理するという特性があり, 単一画像超解像にも応用された．

3　提案手法

本稿では, 劣化処理を施した劣化部分画像 x から復元部分画像 $\hat{y}$ を生成することで, データの高品質化を行う (図 1)．そこで, 部分画像を復元するために用いるための機械学習モデルを提案する．

Copyright is held by the author(s).
The article has been published without reviewing.

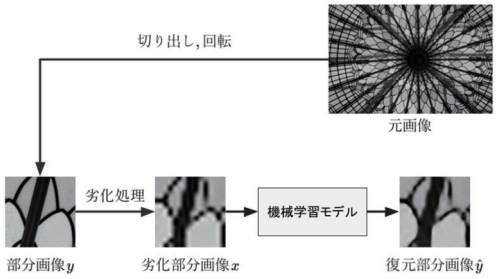


図 1　手順概説

3.1　高品質化の手順

元画像のデータセットからランダムに画像を選出して 128 × 128 × 3, 画素値 255 のカラー部分画像をランダムに切り出し,0 度,90 度,180 度,270 度のランダムで回転を加えたものを部分画像 y とした. 部分画像 y にノイズの追加, 解像度の変更などの任意の劣化処理を行ったものを劣化部分画像 x とした. 劣化部分画像 x から機械学習で部分画像 y を推測して生成した画像を復元部分画像 $\hat{y}$ とした. 劣化部分画像 x と部分画像 y の比較評価には PSNR(Peak Signal-to-Noise Ratio) と SSIM[11] を使用した. 学習に使用する部分画像の枚数 (サンプル数 n) を変更することで, 少数学習が提案する機械学習モデルに及ぼす影響を調査した. 本研究では, 以下 2 つの劣化処理其々に対して, 劣化部分画像を作成した.

3.1.1　超解像

倍の単一画像超解像を行った. 部分画像 y の解像度を 16 × 16 × 3 まで縮小したのち Bicubic 法で 128 × 128 × 3 に拡大したものを劣化部分画像 x とした.

3.1.2　ガウシアンノイズ

部分画像 y の各画素, 各色に 8 対して 平均 $\mu = 0$, 標準偏差 $\sigma = 64$(画素値の 1/4) のガウシアンノイズを加えたものを劣化部分画像 x とした.

3.2　提案する機械学習モデル

本稿では少数学習での過学習を防ぎつつ, 複数の高品質化のアプローチを実現する機械学習モデルを提案する (図 2). 図において, 左の U-Net 構造, 中央の SRCNN 構造, 右の迂回路によって構成した. Loss 関数は平均二乗法 (MSE) を使用した. 最終層以外の各層の活性化関数は ReLU を使用し, 最終層は活性化関数無しとした.

3.2.1　U-Net

U-Net は全体的特徴と局所的特徴の両方を結合する構造を持つ (図 3). U-Net の構造を超解像に利用するに際し, 性能に差異が認められなかったため, 計算コスト削減のために結合部分を加算に変更した.

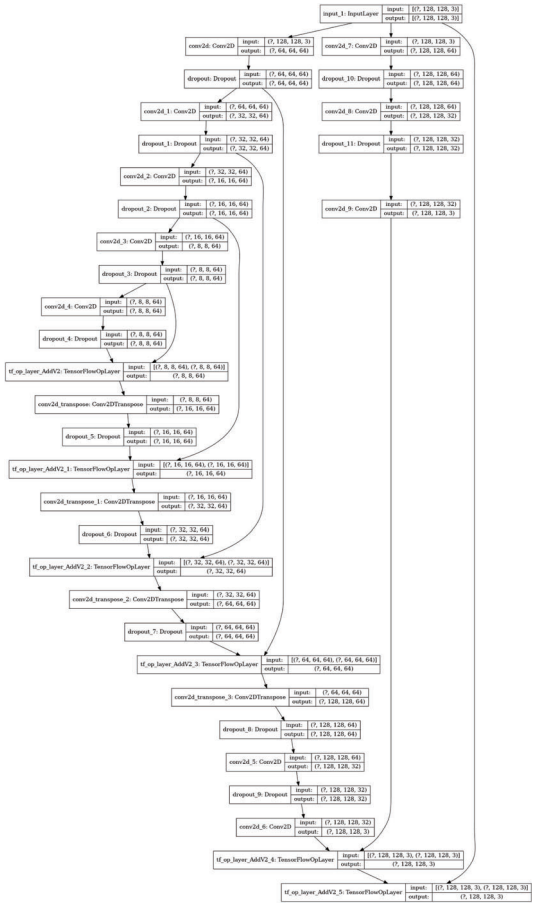


図 2 提案モデルのネットワーク構造

3.2.2　SRCNN

提案する SRCNN 構造は 3 層の CNN によって構成した. U-Net と比較した際, 超解像で劣りガウシアンノイズ除去に優れていたため本モデルに使用した.

3.2.3　迂回路

Dropout による過学習対策は, 精度とトレードオフの関係にある. そこで Kaiming,H. et al.[12] の提案した ResNet を参考にして劣化部分画像 x と復元部分画像 $\hat{y}$ に直通の迂回路を設けた. 迂回路を設けることで, 機械学習モデルの積層や Dropout による元画像の情報の劣化を防ぐことが期待できる.

4　実験方法

4.1　使用データ

学習と評価には高品質画像データセットである DIV2K を使用した [13].DIV2K は学習用 800 枚, 評価用 100 枚の様々なジャンルの画像によって構成されており, 主に単一画像超解像の評価に利用されている. 機械学習に使用するための部分画像のサンプルを学習用 10000 枚, 評価用 1000 枚生成した.

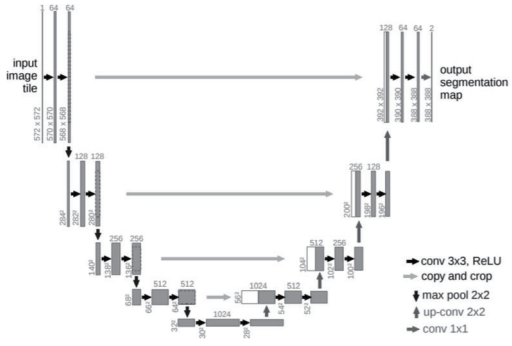


図 3　U-Net のネットワーク構造 [Ronneberger, O., et al. 2015]

4.2　学習パラメータ

最適化には Adam をパラメータを $\beta_1 = 0.1, \beta_2 = 0.999$, 学習率 0.0005 で使用した [14]. 過学習を抑止するための Dropout 率は 0.2 とした. トレーニングの学習方式は, オンライン学習 (batch size=1) で行った. 学習に使用する部分画像のサンプル数 n は [10000,1000,100] の 3 パターンで行った. 学習の backpropagations 数は 100000 とした.

4.3　環境

学習データセット作成などの画像処理に OpenCV を使用した. 深層学習フレームワークは Tensorflow, GPU は NVIDIA RTX2080 を使用した. PSNR と SSIM による評価の為に Scikit-image を使用した.

5　実験結果

5.1　ガウシアンノイズ除去による SRCNN の構造の評価

ガウシアンノイズ除去性能を向上させる目的で設けた SRCNN 構造の有効性を実験した結果を示す (図 4). 提案する機械学習モデルにおいて,PSNR,SSIM 両観点から,SRCNN 構造がガウシアンノイズ除去に有効であることが判明した.

5.2　超解像による迂回路の評価

Dropout による情報劣化を低減する目的で設けた迂回路の有効性を実験した結果を示す (図 5). PSNR,SSIM 両観点から迂回路が存在しないと, 元の画像より劣化することが判明した. また, 迂回路が存在しない場合は色彩が暗くなる傾向があった.

超解像の性能は, ガウシアンノイズと比較して学習に使用するサンプル数 n に大きく依存することが判明した. 具体的には, サンプル数 n が 100 未満になると過学習による情報劣化のリスクが上昇すると考えられる.

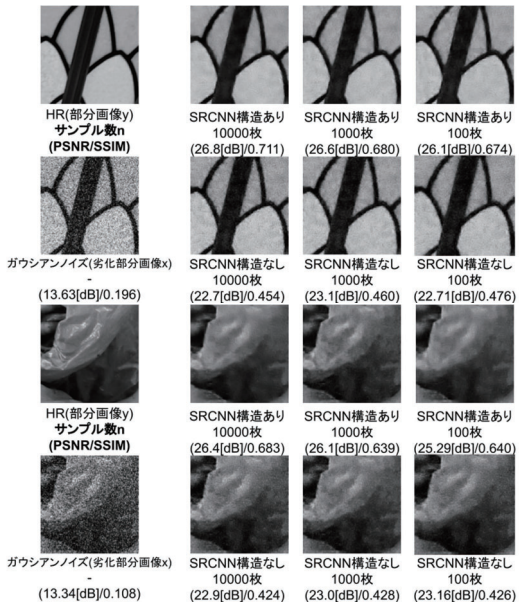


図 4 SRCNN 構造の評価

6　考察

本稿では, 様々なケースに対応した 2 次元データ高品質化のために, 複数の機械学習モデルを組み合わせる手法を提案した. 提案した機械学習により, 深層学習のネットワーク構造において, 複数の目的のネットワークを並列に組み合わせることにより, 有効性を示せるタスクの幅を拡張可能であることを示した. 今後の検討課題として, 二次元データ高品質化のために, ボケや歪みなどのより幅広いタスクに対応する必要があると考えられる. また, 画像以外の二次元データの高品質化にも対応する必要があると考えられる. そのため, 今後は対応するタスクや状況の設定と, 機械学習モデルの機能拡張を検討する.

謝辞

本研究は，JSPS 科研費 20K12081，野村マネジメントスクール研究助成及び東京都立大学傾斜的研究費 (全学分) 学長裁量枠国際研究環支援による.

参考文献

[1] 小野寺康祐, 井上 博夏, 山本 光生ほか: 機械学習による月面 DEM の高解像化, 宇宙航空研究開発機構研究開発報告, Vol.9, No.1,pp. 22-32, 2020.

[2] 伊藤喜代志: 機械学習による超解像技術を活用した詳細な深海海底地形図の作成, 日本水産工学会誌 , Vol.56, No.1, p47-50, 2019

[3] Wei, S., Wu, W., Jeon, G., et al.: Improving resolution of medical images with deep dense convolutional neural network, Concurrency and Computation: Practice and Experience, Vol.32, No.1, e5084., 2020.

[4] Manifold, B., Thomas, E., Francis, A. T., et al.: De-

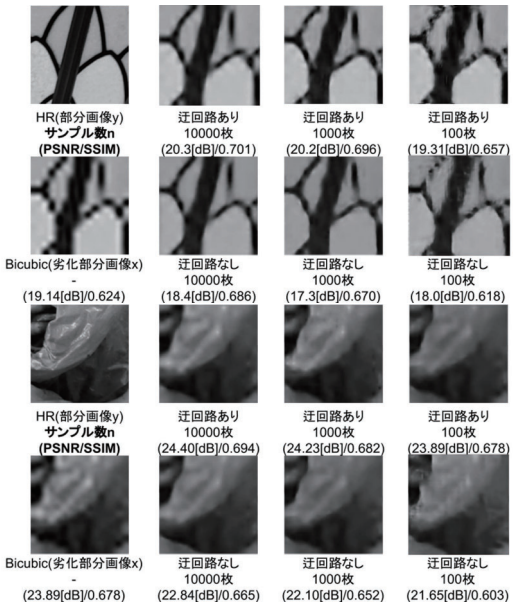


図 5 迂回路の評価

noising of stimulated Raman scattering microscopy images via deep learning,, Biomedical optics express, vol.10, No.8, 3860-3874, 2019.

[5] 小松里奈: U-Net による手書き文字画像内のノイズ除去, 人工知能学会全国大会論文集, Vol.32, No.1, p.4M101-4M101, 2018.

[6] Samuel ,A. L. : Some studies in machine learning using the game of checkers, IBM Journal of research and development, Vol.3, No.3, pp. 210-229, 1959.

[7] Lecun, Y., Bottou ,L., Bengio, Y., et al.: Gradient-based learning applied to document recognition, Proceedings of the IEEE, Vol.86, No.11, pp. 2278-2324, 1998.

[8] Srivastava, N., Hinton, G., Krizhevsky, A., et al.: Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research, Vol.15, No.1, pp. 1929-1958, 2014.

[9] Dong ,C., Loy, C. C., He, K., Tang, X., et al.: Learning a Deep Convolutional Network for Image Super-Resolution, in Proceedings of European Conference on Computer Vision (ECCV),pp. 184-199, 2014.

[10] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation, International Conference on Medical image computing and computer-assisted intervention (MICCAI), pp. 234-241, 2015.

[11] Z. Wang, A. C. Bovik, H. R. Sheikh, Simoncelli, E. P.:Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4):600-612. 2004

[12] He, K., Zhang, X., Ren, S., et al. : Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[13] Timofte, R., Agustsson, E.,Gool, L. V., et al.: NTIRE 2017 challenge on single image super-resolution: Methods and results, IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

[14] Kingma, D. P., Ba, J. L. , et al.: Adam: A Method for Stochastic Optimization, International Conference on Learning Representations (ICLR), 2015.

ARG WI2 No.16, 2020
WI2-2020-15

SNSを用いたトレンドスポットの検出の検討

中田 朋寛[†,a] 三浦 拓也[†,b] 宮坂 和希[†,c] 荒木 徹也[††,d]
遠藤 雅樹[†††,e] 土田 正士[†,f] 山根 康男[†,g] 平手 守浩[††††,h]
眞浦 雅夫[††††,i] 石川 博[†,j]

† 東京都立大学大学院システムデザイン研究科 †† 群馬大学理工学部 ††† 職業能力開発総合大学校 ††††
アイシン・エィ・ダブリュ株式会社 CSS 本部コネクティッドソリューション部

a) *nakata-tomohiro@ed.tmu.ac.jp* b) *miura-takuya1@ed.tmu.ac.jp*
c) *miyasaka-kazuki@ed.tmu.ac.jp* d) *tetsuya.araki@gunma-u.ac.jp*
e) *endou@uitec.ac.jp* f) *tsuchida@tmu.ac.jp* g) *yamane@tmu.ac.jp* h) *i24841_hirate@aisin-aw.co.jp*
i) *i45588_maura@aisin-aw.co.jp* j) *ishikawa-hiroshi@tmu.ac.jp*

概要 SNS では，日々多くのユーザによって実世界での出来事が大量に投稿されている．そのため，実世界で起きているイベントやトレンドなどの影響により，しばしば平常時の投稿数と比べ， 大きく増加するバーストと呼ばれる現象が観測される．SNS 内のトレンドは一般に公開されているが，トレンドの地域性やトレンドの場所（トレンドスポット）を知ることは難しい．そこで本研究では，SNS の一つである Twitter を用いて地域のトレンドワードとトレンドスポットを検出する手法の提案を行う．具体的には以下の 4 段階の検出を行う．
(1) トレンドワード候補検出 (2) トレンドワード検出 (3) トレンドスポット候補検出 (4) トレンドスポット検出
名古屋市の Twitter データを用いてトレンドスポットの検出を行い，提案手法の有効性を検証した．本研究により，SNS のトレンドをナビゲーションなどのサービスへの応用の可能性を示した．

キーワード Twitter，ハッシュタグ，イベント情報抽出，観光

1 はじめに

近年，Social Networking Service（以下 SNS）などのソーシャルメディアが普及してきている．SNS には，ユーザの実世界の情報が投稿されることが多く，その内容はユーザの体験，意見，感想など様々なものがある．SNS の一つである Twitter は，tweet と呼ばれる短文を時間・場所問わず気軽に投稿することができ，他ユーザに拡散されやすいという特徴がある．Twitter ではハッシュタグと呼ばれるタグをユーザが自身の tweet に付与することができる．ハッシュタグは tweet 内容のキーワードのようなもので,「# word」のように「#」を先頭に表わされる．基本的に同じハッシュタグが使用されている tweet 集合は同じカテゴリや属性に対し言及しており，任意の期間におけるその集合の大きさは実世界のイベントに影響される．そのため，Twitter ではしばしば任意のハッシュタグを含む tweet 数が平常時と比べ，大きく増加するバーストと呼ばれる現象が観測される．その原因の一つに，社会での流行している物・事象があり，一般的にその物・事象はトレンドと呼ばれる．トレンドは，Twitter 内で公開されており，どんな内容・話題が流行っているのかをリアルタイムに知ることができる．そのため，Twitter は実世界のトレンドを知る手段として用いられることが多く，その情報から社会のニーズやユーザの興味関心など様々な情報が分かる．その中には，場所についての内容も稀に見られ，それらに付随する情報を旅行の参考情報として利用する人もいる．しかし，トレンドの中には物・事象など様々なものがあり，トレンドである場所（以下，トレンドスポット）のみを知るのは難しい．また，Twitter からトレンドスポットの位置を取得することはできず，Web サイトでの検索を強いられるためユーザの負担が増えてしまう．

そこで本研究では，Twitter でのハッシュタグの地域内でのバーストに着目し，トレンドスポットを検出する手法を提案する．提案手法は，指数平滑移動平均を用いてバーストしているハッシュタグをトレンドワードとして検出し，各トレンドワードについて関連するスポットをトレンドスポットとして検出する．また本研究では，地域内でのバーストしているハッシュタグを検出するため，地域内でのトレンドスポットの検出を可能にしている．そのため，Twitter での取得が難しかった地域でのトレンドワードが検出でき，ユーザが取得できる有効な参考情報量の向上が期待できる．本論文は，以下の構成に従う．2 章では，本研究に関連する研究について述べる．3 章では，提案手法について述べる．4 章では，実験結果とそれに対しての考察について述べる．5 章では，本論文のまとめと今後の課題について述べる．

Copyright is held by the author(s).
The article has been published without reviewing.

2　関連研究

本章では関連研究として，Twitter を用いた時空間イベント検出に関する研究と Twitter でのハッシュタグに関する研究について述べる．

2.1　Twitter を用いたイベント検出に関する研究

Twitter を用いたイベント検出に関する多くの研究が行われている．古澤ら [1] は特定地域名を含む tweet を教師なし学習を行うことでリアルタイムに起きているイベントを抽出している．特定の地名と関連のあるイベント単語のみの抽出を目的としており，最終的に POI の抽出を目的としている本研究とは本質的に異なる．

山田ら [2] は SVM と CRF を用い，Twitter からイベント情報を自動的に抽出する技術を提案している．Twitter で投稿されたイベントの告知ツイートやイベント名称の特徴に着目している点で本研究とは異なる．

中澤ら [3] は特定エリア内の多数の tweet が投稿されるようなイベント地点を検出し，その場所でのイベント内容を推定している．任意の地域の POI を検出するのではなく特定の場所でのイベントを検出している点で本研究とは異なる．

MATHIOUDAKIS ら [6] は任意のキーワードの出現頻度に着目し，バーストしている単語を検出している．また，検出された単語をそれらの単語の共起によりグループ化している．しかし，トレンドそのものの順位付けや分析を目的としている点で本研究とは異なる．

LEE ら [7] は任意の地域の特徴やその地域でのイベントを SNS のユーザの行動から検出している．単語のバーストからイベントを検出しているのではなく，任意の地域内の tweet から得られるユーザの行動から検出している点で本研究とは異なる．

2.2　ハッシュタグに関する研究

本岡ら [4] は Twitter のハッシュタグを用いて，類似するイベントを発見している．ハッシュタグを認知度によりハッシュタグをフィルタリングし，類似度で評価することで入力ハッシュタグと類似するイベントを検出している．しかし，ハッシュタグ間の類似性を利用し類似イベントを検出しているが，ハッシュタグのバーストに言及していない点で本研究とは異なる．

福山ら [8] はバーストしたハッシュタグクラスタリングし，バーストしていないハッシュタグに割り当てる手法を提案している．バーストしていないハッシュタグに着目している点で本研究とは異なる

3　提案手法

本章では，本研究のトレンドスポット検出手法について述べる．トレンドスポット検出の流れを図 1 に示す．まず，本研究でのトレンドワード及びトレンドスポットの定義について述べ，その後提案手法について述べる．

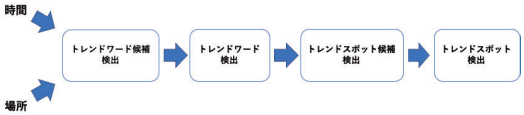


図 1 トレンドスポット検出の流れ

3.1　トレンドワード・トレンドスポットの定義

本節では，本研究で使われるトレンドワード及びトレンドスポットの定義について述べる．以下の条件を満たすワードをトレンドワードと定義する．

- 任意の地域で短期間に局所的に投稿ユーザ数が増加しているワード

また，以下の 2 つの条件を満たすワードをトレンドスポットと定義する．

- トレンドワードとの関連性が高いワード
- トレンドワードが検出された地域内に存在するスポット名を指すワード

3.2　提案手法

本節では，ハッシュタグを用いたトレンドスポット検出の 4 段階の処理について述べる．また，過去の実際のトレンドスポットを検出目標スポットとしそのスポットがトレンドであった時間，場所で実験を行い評価するため，トレンドスポット検出への入力は時間，場所の 2 値と，その場所のスポットのデータセットとする．

3.2.1　トレンドワード候補検出

本節では，トレンドスポット検出の第 1 段階目の処理であるトレンドワード候補検出について述べる．以下にトレンドワード候補検出の処理手順を示す．

(1) 本研究で使用する位置情報付き tweet データから入力された場所，入力された時間から過去 t_1 日間で投稿された tweet を取得

(2)(1) で取得した tweet 内に含まれるハッシュタグを取り出し，ハッシュタグごとに投稿数を集計

(3) 投稿数が 2 件以上のハッシュタグをトレンドワード候補として検出

(1) の処理の定数 t_1 は，トレンドワード候補とする期間（以下，トレンドワード候補期間）を表す定数とする．(3) の処理はユーザが独自に付与した一般的でないハッシュタグの検出を防ぐための処理である．

3.2.2　トレンドワード検出

本節では，トレンドスポット検出の第 2 段階目の処理であるトレンドワード検出について述べる．トレンド

ワード候補検出では，3.2.1 節で検出したトレンドワード候補それぞれに対し，以下の処理手順を行う．
(1) 入力された時間の過去 t_2 日間のトレンドワード候補を含む tweet を取得
(2) (1) で取得した tweet を投稿ユーザ数で日別に集計
(3) ユーザ投稿数推移から 1 日ごとの指数平滑移動平均 [9] の値 (EMA(n)) を算出
(4) 入力された日の投稿ユーザ数が, その日の指数平滑移動平均の値より多い場合， そのトレンドワード候補をトレンドワードとして検出
(1) の処理の定数 t_2 は，指数平滑移動平均を計算する期間（以下，指数平滑移動平均計算期間）を表す定数とする，(2) の処理では，同ユーザが短期間に投稿した同じハッシュタグを含む tweet による投稿数の偏りを防ぐため，投稿ユーザ数で集計している．(3)，(4) の処理では，過去の投稿ユーザ数の推移から局所的に増加しているワードを検出するため，移動平均を用いる．また単純移動平均と比べ，直近の推移に比重を置き，データに対し高い感度が得られるため指数平滑移動平均を用いる．EMA(n) を n 日目の指数平滑移動平均の値，C を指数平滑係数，S を単純移動平均の期間，y を n 日目の出現回数とすると， C を求める式と指数平滑移動平均の値を求める式はそれぞれ式 (A)，式 (B) のように表される．

$$C = \frac{2}{S-1} \qquad (A)$$

$$EMA(n) = C * y + (1 - C * y) * EMA(n-1) \quad (B)$$

3.2.3 トレンドスポット候補検出

本節では，トレンドスポット検出の第 3 段階目の処理であるトレンドスポット候補検出について述べる．トレンドスポット候補検出では，3.2.2 節で検出したトレンドワードそれぞれに対し，以下の処理手順を行う．
(1) 入力された時間の過去 t_3 日間のトレンドワードを含む tweet を取得
(2) (1) で取得した tweet 内に含まれるハッシュタグを全て取得
(3) (2) で取得したハッシュタグの内， 入力された場所に存在するスポット名と一致するハッシュタグをトレンドスポット候補として検出
(1) の処理の定数 t_3 は，トレンドスポット候補を検出する期間（以下，トレンドスポット候補検出期間）を表す定数とする．(2) の処理では，トレンドワードと関連のあるワードを検出する．本研究では，トレンドワードをハッシュタグに含むトレンドスポット候補検出対象期間内の tweet に存在するハッシュタグを全て関連ワードとしている．(3) の処理では，(2) の処理で取得したハッシュタグそれぞれに対し， 入力された場所に存在するスポット名と完全一致するかどうかを判別し，トレンドスポット候補を検出する．

3.2.4 トレンドスポット検出

本節では，トレンドスポット検出の第 4 段階目の処理であるトレンドスポット検出について述べる．トレンドワード候補の中には，トレンドワードとの関連性の低いものや，ある特定のユーザのみが使用しているハッシュタグの組み合わせが存在し，それらは誤検出に繋がる．そのためトレンドスポット検出では，3.2.3 節で検出したトレンドスポット候補それぞれに対し， トレンドワードとの類似度という尺度を用いた評価を行い，トレンドスポットを検出する．類似度は，トレンドワードとトレンドスポット候補それぞれを投稿したユーザの集合から，それらのワードがどれだけ類似しているかを判定する尺度である．類似性が高く一般的に使用されるハッシュタグの組み合わせはハッシュタグだけでなく，テキストにも共通して出現する可能性が高い．そこで本研究では，テキストを用いた類似度とハッシュタグを用いた類似度の差を用いた評価を行い， その値が閾値 th 以上のトレンドスポット候補をトレンドスポットとして検出する．任意のワード w をハッシュタグに含む t_4 日間（以下，類似度計算期間）の tweet の集合 $T_{hash}(w)$ の中で，ワード w' をハッシュタグに含む確率を類似度 $simh(w, w')$ と定義する．任意のワード w をハッシュタグに含む tweet の集合の大きさを $|T_{hash}(w)|$，その中でワード w' をハッシュタグに含む tweet の集合の大きさを $|T_{hash}(w' \mid w)|$ とすると，類似度 $simh(w, w')$ の式は次の (C) 式のように表される．

$$simh(w, w') = \frac{|T_{hash}(w' \mid w)|}{|T_{hash}(w)|} \qquad (C)$$

同様にテキストを用いた類似度を定義する．任意のワード w をテキストに含む t_4 日間の tweet の集合を $T_{text}(w)$，ワード w' をテキストに含む確率を類似度 $simt(w, w')$, 任意のワード w をテキストに含む tweet の集合の大きさを $|T_{text}(w)|$，その中でワード w' をテキストに含む tweet の集合の大きさを $|T_{text}(w' \mid w)|$ とする．類似度 $simt(w, w')$ の式はの (D) 式のように表される．

$$simt(w, w') = \frac{|T_{text}(w' \mid w)|}{|T_{text}(w)|} \qquad (D)$$

評価式は次の (E) 式のように表される．

$$simt(w, w') - simh(w, w') >= th \qquad (E)$$

以上の式 (E) を満たすトレンドスポット候補をトレンドスポットとして検出する．

表 1　実験条件

検出目標スポット	別小江神社
実験時間	2019 年 10 月 22 日 12 時
実験地域	愛知県名古屋市

4　実験

本章では提案手法の有効性を確認するため， 実際に Twitter から収集したツイートを用いて，3 章で提案した手法によりトレンドスポット検出を行う．本研究では，検出したスポットがトレンドであるかの判断ができないため，実世界で実際にトレンドであったスポット (以下， 検出目標スポット) を予め用意し，実験を行う．検出目標スポットが実世界でトレンドであった時間，そのスポットが存在する場所を入力とし，実験を行う．

4.1　検出目標スポットの決定

本節では，実験で用いる検出目標スポットについて述べる．実験条件を表 1 に示す．

表 1 で示される別小江神社は，愛知県名古屋市にある御朱印が有名な神社である．しかし，Twitter での投稿ユーザ数は 2018 年 10 月 22 日から 2019 年 10 月 22 日の 1 年間で 53 人と非常に少ない．本研究では地域性が高いスポットや投稿 tweet 数の少ないスポットへの適用も考慮するため，検出目標スポットを別小江神社とする．また，2019 年 10 月 22 日は即位礼聖典の儀が行われた日であり，別小江神社では記念の御朱印販売などイベントが行われていた．そのため，本研究では 2019 年 10 月 22 日 12 時を実験時間とする．

4.2　データセット

4.2.1　Twitter のデータセット

本節では，本研究の実験に使用した Twitter のデータセットについて述べる．2018 年 1 月から 2020 年 6 月までの位置情報付き tweet をランダムに取得した．その結果，255,509,636 件のデータが集まった．その中で，実験で使用する愛知県名古屋市のデータは 5,131,208 件であった．

4.2.2　スポットのデータセット

本節では，実験で使用する愛知県名古屋市のスポットのデータセットについて述べる．本研究では，Yahoo ローカルサーチ API を用いて愛知県名古屋市のスポットデータを取得した．そのスポットデータにはスポット名とそのスポットのカテゴリ，そのスポットが存在する緯度経度が存在する．スポットのカテゴリは，グルメ，ショッピング，レジャーエンタメ，暮らし生活の 4 つである．その結果，それぞれのカテゴリで 15,525 件，9,693 件，2,714 件，30,899 件の全 58,831 件のスポットデータを取得できた．

表 2 実験パラメータ

パラメータ	意味	値
t_1	トレンドワード候補期間	3
t_2	指数平滑移動平均計算期間	30
t_3	トレンドスポット候補検出期間	60
t_4	類似度計算期間	365
S	単純移動平均の期間	7
th	評価値の閾値	0

表 3 トレンドワード候補検出結果

RWC2019	名古屋まつり	本田圭佑
ペプシ	ジャパンコーラ	ミラティブ
AQUOSR2	大須大道町	菊花賞
名古屋グランパス	名古屋城	synth
大須	台風 20 号	他 581 件

4.3　実験パラメータ

本節では，実験で用いたパラメータについて述べる．実験パラメータを表 2 に示す．

4.4　結果

本節では，実験の結果について述べる．4.1 節，4.2 節，4.3 節で示した条件の下， 3 章で提案した検出を行った結果を示す．トレンドワード候補検出の結果の一部を表 3 に示す．

4.4.1　トレンドワード候補検出結果

本節は提案手法の第 1 段階の処理であるトレンドワード候補検出の結果及びその結果に対する考察を述べる．トレンドワード候補検出の結果の一部を表 3 に示す．

トレンドワード候補の検出では,「RWC2019」や「台風 20 号」は 2019 年 10 月に起きたイベントのことを指すワードや， 「名古屋城」や」「大須」などの地名を指すワードなど全 595 ワードが検出された．中には,「本田圭佑」や「ペプシ」など Twitter でのツイートキャンペーンで投稿されたワードや， 「synth」など一見何を指すワードなのかわからないワードも多く検出された．

4.4.2　トレンドワード検出結果

本節は提案手法の第 2 段階の処理であるトレンドワード検出の結果及びその結果に対する考察を述べる．トレンドワード検出の結果の一部を表 4 に示す．

トレンドワードの検出では,「台風 20 号」や「即位礼正殿の」など 2019 年 10 月に起きたイベントのことを指すワードや， 「新幹線」や「地下鉄」などの一般名詞など全 39 ワードが検出された．中でも,「プレモル」や「サッポロ一番 W」など Twitter でのツイートキャンペーンで投稿されたワードが多く検出された．また，ト

表 4 トレンドワード検出結果

天気	台風２０号	即位礼正殿の儀
地下鉄	女子大	オカマ
アイラブ塩	ハイモニ	御朱印
ジョーカー	台湾ラーメン	プレモル
サッポロ一番 W	新幹線	他 25 件

表 5 トレンドスポット候補検出結果

トレンドスポット候補	関連するトレンドワード
味仙矢場店	台湾ラーメン
名古屋めし	即位礼正殿の儀
名古屋城	天気
日の出	新幹線
葵	女装
葵	女装
名古屋市立大学	地下鉄, 女子大
伊奴神社	御朱印
金山	地下鉄
幸	御朱印
幸	御朱印
ひかり	地下鉄, 新幹線
テレビ塔	台湾ラーメン
ウニ	大阪
東京	新幹線
別小江神社	御朱印
三輪神社	御朱印

レンドワードの中には指数平滑移動平均の値が 1 以下のものが多く，1 件のみの投稿で検出されるものが多く見られた．

4.4.3 トレンドスポット候補検出結果

本節は提案手法の第 3 段階の処理であるトレンドワード検出の結果及びその結果に対する考察を述べる．トレンドスポット候補検出の結果を表 5 に示す．

トレンドスポット候補の検出では，神社や飲食店が多く検出された.「葵」と「幸」は同音で異なる店舗であるため，それぞれ 2 スポット検出された．トレンドスポット候補と関連するトレンドワードを比較すると,「名古屋城」と「天気」，「名古屋市立大学」と「女子大」など一見関連性の低い組み合わせのものが多く見られた．また,「東京」や「ウニ」など日常的に使われる言葉と同じ名前のスポットも多く検出された．実際に投稿された tweet 中には，名古屋市のスポットである「東京」を指している内容のものは存在しなく，そのようなスポットは誤検出されている可能性が高いと考えられる．トレンドワードと関連するワードの内，トレンドスポット候補に検出されなかったスポットの中に「味仙今池本店」や「大須商店街」,「白龍神社」などスポット名と考えられるワードがいくつか存在した．このようなワードは，スポットのデータセット内に含まれていないか，スポットのデータセットに登録されているスポット名と Twitter にハッシュタグとして投稿されるワードの表記方法が異なることが考えられる．

表 6 トレンドスポット検出結果

トレンドスポット名	スポットのカテゴリ
味仙矢場店	グルメ
名古屋めし	ショッピング
金山	グルメ
別小江神社	暮らし生活
三輪神社	暮らし生活

4.4.4 トレンドスポット検出結果

本節は提案手法の第 4 段階の処理であるトレンドスポット検出の結果及びその結果に対する考察を述べる．トレンドスポット検出の結果とそれぞれのスポットのカテゴリを表 6 に示す．

トレンドスポットの検出では，暮らし生活カテゴリのスポットが 2 スポット，グルメカテゴリのスポットが 2 スポット，ショッピングカテゴリのスポットが 1 スポット検出された．また，検出目標スポットである別小江神社を検出することでき，手法の有効性を確認できた.「名古屋めし」はショッピングの中の通信販売カテゴリに属している名古屋市のスポットであるが，tweet で使用されている「名古屋めし」は名古屋市にある飲食店全般を表すワードして用いられている．また,「金山」は名古屋市の飲食店であるが，tweet で使用されている「金山」は名古屋市にある地下鉄の駅名またはその近辺の地域を表すワードして用いられている．そのため,「名古屋めし」と「金山」は同音異義語により誤検出されていると考えられる．

5 まとめ

本研究では，4 段階の検出を用いることで Twitter のハッシュタグからトレンドであるスポットを検出する手法を提案した．そして，愛知県名古屋市にある別小江神社を検出目標スポットとして実験を行い，提案手法の有効性を確認した．

今後の課題として，キャンペーンの tweet や店の広告や宣伝の tweet などの影響でその投稿に含まれるスポット名を多く検出してしまうため，そのような投稿を除去することが挙げられる．次に，東京や金山など同音異義語による誤検出が多く見られた．そのようなワードを

含む投稿を確認しどの意味，どのスポットとして投稿中で使用されているのかを判断する必要がある．また，スポットのデータセットに含まれるスポット名と Twitter で投稿されるスポット名の表記揺れが原因で検出できなかったスポットがいくつか存在するため，表記揺れの対策をする必要がある．さらに，指数平滑移動平均の値が 1 以下であるため，投稿数が一件前後の極端に投稿数が少ないワードが検出されているため，そのようなワードを除去する処理が必要である．

今後の展望として，Twitter だけでなく instagram など他の SNS を検出に用いることなどが挙げられる．これによって，Twitter でのスポット名を含む投稿数の少なさが補え，現状より様々な種類のスポットが検出できると考えられる．また，特徴や SNS での投稿数などが異なる様々な地域での適用も挙げられる．そのため，名古屋市以外での場所での実験を行い，パラメータの設定など様々な地域に応じた検出を考える必要があると考えられる．

参考文献

[1] 古澤康太，秋岡明香．マイクロブログ情報解析によるイベント検出の提案．情報処理学会第 82 回全国大会

[2] 山田渉，菊地悠，落合桂一，鳥居大祐，稲村浩，太田賢．マイクロブログを用いたイベント抽出技術，情報処理学会論文誌　Vol.57，No.1，pp.123-132(2015)．

[3] 中澤昌美，池田和史，服部元，小野智弘．位置情報付きツイートからのイベント検出手法の提案．情報処理学会第 74 回全国大会

[4] 本岡亮，湯本高行，新居学，高橋豊，角谷和俊．Twitter ハッシュタグを用いた類似イベント検索．DEIM Forum 2011 A1-5.

[5] 木村輔，宮森逅. 共起と潜在トピックを考慮したハッシュタグ間関係の分類手法. 電子情報学会論文誌　Vol. J98-D，N0.8，pp.1151-1161.

[6] MATHIOUDAKIS，M.，and N. KOUDAS. 2010. TwitterMonitor. Trend detection over the Twitter stream. In SIGMOD Conference，Indianapolis，pp. 1155–1158.

[7] LEE，R.，and K. SUMIYA. 2010. Measuring geographical regularities of crowd behaviors for Twitter-based geo- social event detection. In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks，LBSN ’10，ACM，New York，NY, pp. 1–10.

[8] 福山怜史，若林啓．バーストを考慮したハッシュタグのクラスタリング手法の提案．情報処理学会研究報告．Vol.2017-DBS-165 No.17．pp.1-7.

[9] Hunter, J. S. (1986)，“The Exponentially Weighted Moving Average,”Journal of Quality Technology，18，203-210.

ARG WI2 No.16, 2020
WI2-2020-16

SNSを用いた短期間イベント分析

宮坂 和希[†,a] 中田 朋寛[†,b] 三浦 拓也[†,c] 荒木 徹也[††,d]
土田 正士[†,e] 山根 康男[†,f] 平手 守浩[†††,g] 眞浦 雅夫[†††,h]
石川 博[†,i]

† 東京都立大学大学院システムデザイン研究科 †† 群馬大学理工学部
††† アイシン・エィ・ダブリュ株式会社 CSS 本部 コネクティッドソリューション部

a) *miyasaka-kazuki@ed.tmu.ac.jp* b) *nakata-tomohiro@ed.tmu.ac.jp* c) *miura-takuya1@ed.tmu.ac.jp*
d) *tetsuya.araki@gunma-u.ac.jp* e) *tsuchida@tmu.ac.jp* f) *yamane@tmu.ac.jp*
g) *i24841_hirate@aisin-aw.co.jp* h) *i45588_maura@aisin-aw.co.jp* i) *ishikawa-hiroshi@tmu.ac.jp*

概要 株価やオリンピックなど不特定多数が関わるイベントに対する世論の反応を予測するとき，SNS の分析を活用することがある．SNS で分析することで比較的低コストにイベントに対する反応を把握できる．しかし，オリンピックなど 1 日以上続くイベントに対する分析は存在するが，1 日未満の短期間イベントに対する分析は現状ほとんど存在しない．短期間イベントにも，2020 年に東京で発生したブルーインパルスや日本中で発生した打ち上げ花火など，国民の関心が非常に高いイベントは存在する．そのため短期間イベントを SNS で分析することは有用である．そこで本研究では代表的な SNS であるツイッターを用いて，前述のブルーインパルスや花火などの短期イベントに関する投稿を感情分析することで SNS 上の世論の分析を行った．本研究の結果により，短期間イベントに対して比較的低コストで世論の反応を把握することが可能になることが期待できる．

キーワード イベント分析，SNS 分析，ツイッター分析，感情分析

1 はじめに

株価やオリンピックなど不特定多数が関わるイベントに対する世論の反応を予測するとき，一般的に SNS を用いて予測する方法が存在する．SNS はオンラインの世界で意見を表明し他の人と交流するための手段の一つであり，その中でもツイッターは非常に人気の SNS である．そのためイベントを SNS で分析するときツイッターは世論の反応を評価するためのソースとしてしばしば用いられる．SNS を用いる他にはアンケートが挙げられるが，これと比較すると SNS の評価の方が比較的低コストにイベントを分析することができるため，現在も SNS によるイベント分析の研究が盛んに行われている．現在行われている研究の中には，オリンピック [1] やワールドカップ [2] など，1 日以上続く大きなイベントに対する世論の反応を SNS で評価する研究も行われている．

しかし，前述のようなオリンピックやワールドカップなどの 1 日以上続く大きなイベントに対する SNS の分析は存在するが，1 日未満の短期間イベントに対する分析は現状ではほとんど存在しない．短期間イベントの中には，2020 年 5 月 29 日に東京で発生した，医療従事者等に対する敬意や感謝を示すためのブルーインパルス飛行[1]や，2020 年 6 月 1 日に全国各地で発生した，花火業者が一斉に花火を打ち上げる打ち上げ花火プロジェクト[2]など，国民の関心が非常に高い短期間イベントが存在する．そのため短期間イベントを分析し世論の反応を評価する研究は需要があると考えられる．

そこで本研究では，前述のブルーインパルスや花火を用いて，1 日未満の短期間イベントに対してツイートを感情分析し，イベントとの相関関係があるか，どのような影響を与えるかの調査を行う．ツイートの感情分析は，感情辞書と Wordnet と絵文字辞書を組み合わせた辞書を作成することで行う．またその辞書に含まれない単語は word2vec を用いて感情を推測する．本研究の結果により，1 日未満の短期間イベントに対して比較的低コストで分析し世論の反応を評価することが可能になることが期待できる．

本論文の構成は次の通りである．2 章では，本研究で利用するデータセットの詳細，またそのデータセットの前処理方法について述べる，3 章では，感情の定義と，本研究で利用する感情分析モデルの詳細について述べる．4 章では，3 章の感情分析モデルを 2 章のデータセットに適用して作成した，感情分析結果を分析し考察を述べる．5 章では，本研究のまとめと今後の課題を述べる．

2 データセット

2 章では，本研究で利用するデータセットの詳細，またそのデータセットの前処理方法について述べる．

Copyright is held by the author(s).
The article has been published without reviewing.

[1] https://www.mod.go.jp/asdf/news/houdou/R2/20200529.pdf

[2] https://www.cheeruphanabi.com/

2.1 実験で用いるイベント

本研究では１日未満の短期間イベントとして，2020 年に発生したブルーインパルスと打ち上げ花火の 2 つに対して実験を行う．この２つのイベントの詳細を以下に示す．

- 2020 年 5 月 29 日 12:40-13:00 に行われたブルーインパルス飛行．新型コロナウイルス感染症へ対応中の医療従事者等に対する敬意や感謝を示すために東京都上空を 2 度周回した．周回ルートはイベント開始２時間前に事前に告知していた．
- 2020 年 6 月 1 日 21:00-21:05 に行われた打ち上げ花火プロジェクト．コロナの収束と励ましを込めて，花火業者が一斉に花火を打ち上げた．時間の告知は事前にしていたが，人が密集しないように場所の告知はしなかった．

どちらも 1 日未満の短期間イベントであり，かつ非常に世論の関心が高いイベントであるため，本研究の短期間イベントに適していると考えられる．

2.2 前処理方法

本研究では，代表的な SNS であるツイッターをデータセットとして用いて実験を行う．データセットの前処理方法を以下に示す．

まず初めに，短期間イベントの分析を行うため，本実験で使用するイベントと関連のあるツイートを抽出する．抽出条件として以下の処理を行う．

- イベントは 2020 年 5 月 29 日と 2020 年 6 月 1 日にそれぞれ発生しているので，2020 年 5 月-2020 年 6 月の間のツイートを取得．
- ブルーインパルスは東京で発生しており，また打ち上げ花火も最もツイート数が多かったのは東京であるため，ツイートに付随している場所の情報から東京のツイートを抽出．
- 表 1 に示したブルーインパルスと打ち上げ花火に関連のあるキーワードを用いて，その単語がツイートテキストもしくはハッシュタグに存在していたらそのツイートを抽出．

抽出の結果，ブルーインパルスに関連するツイート数は 2770, 打ち上げ花火に関連するツイート数は 1026 だった．

次に，感情分析を行うために，形態素解析を行うことで文章を単語のリストに分解した．本研究ではそのための方法として Mecab[3] と Cabocha[4] を用いた．Mecab とはオープンソース形態素解析エンジンで，これを用いることでツイートに対して形態素解析を行うことができる．また Cabocha とは SupportVectorMachines に基づく日本語係り受け解析器で，これを用いることでツイート文章で使われる単語の係り受けを解析することができる．本実験の形態素解析の手順は，ハイパーリンク，助詞，改行，リプライは感情分析の精度を落とすと考えられている [3][4] ため本実験ではツイートから削除した．その後 Mecab と Cabocha を用いることで，イベントに関連するツイートに対して形態素解析および係り受け解析を行った．その結果，ツイート文章を単語リストにしたデータおよび係り受け解析リストを得た．

表 1 Tweet 抽出キーワード

ブルーインパルス	blueimpulse, ブルーインパルスなど計 19
打ち上げ花火	花火，cheeruphanabi など計 12

3 感情分析

この章では 2 章で作成したツイートに対して行う感情分析について述べる．

3.1 定義

感情を定義するために「Plutchik の感情の輪」を用いる [5]．これは基本感情が，「喜び・信頼・恐れ・驚き・悲しみ・嫌悪・怒り・期待」の８つで構成された感情モデルであり，感情を定量化する際に使用されるモデルである．本実験では，このモデルを用いて感情を 8 つに分類する．

3.2 感情辞書

感情分析のために本実験では感情辞書を用いる．感情辞書とは単語一語に対して一つもしくは複数の感情が紐づいている単語リストである．本実験では感情辞書は長岡技術科学大学自然言語処理研究室の感情辞書を用いる[5]．この感情辞書は単語が 48 の感情で分類されているため，Plutichik の感情の輪を用いるために 48 の感情を 8 の感情にクラスタリングした．方法としては，5 人によるアンケート結果から多数決により決定した．クラスタリング結果は表 2 に示す．

また別の感情辞書も用いる．上述の感情辞書のワード数は 1914 であり，語彙数が少ないため感情分析が正しくできない可能性がある．そのため別の感情辞書を作成することで感情辞書を拡張する．方法としては日本語 Wordnet[6] を用いることで Wordnet 感情辞書を作成した．Wordnet とは語を類義関係のセットでグループ化

[3] https://taku910.github.io/mecab/
[4] https://taku910.github.io/cabocha/
[5] http://www.jnlp.org/kamiwaki/gan-qing-ci-shu
[6] http://compling.hss.ntu.edu.sg/wnja/

表 2 感情辞書の 8 感情へのクラスタリング分類結果

喜び	楽しさ，気持ちが良い，誇らしい，喜び，幸福感，興奮，祝う気持ち
信頼	安らぎ，親しみ，感謝，好き，穏やか，尊敬・尊さ
恐れ	不安，恥ずかしい，焦り，恐怖，悩み，心配，恐れ (恐縮等の意味で)，ためらい，緊張
驚き	感動，驚き，困惑
悲しみ	悲しさ，寂しさ，切なさ，苦しさ，失望あわれみ，謝罪，残念，情けない
嫌悪	憂鬱，辛さ，嫌悪，見下し，不快，怠さ，あきれ，きまずさ
怒り	不満，怒り，恨み，悔しさ，妬み，憎い
期待	願望

し，さらに各セットを上位下位関係などの多様な関係で結んだ大規模な意味辞書であり．これを用いることで，単語の類義語を検索することができる．そのため，長岡技術科学大学感情辞書に載っている単語を検索し，類義語を新たな語彙として追加することで感情辞書を拡張した．結果として拡張した感情辞書の語彙数は 6837 となった．

またさらに別の辞書として絵文字キーワード対応辞書[7]を活用した．絵文字キーワード対応辞書とは一つの絵文字に対してその絵文字に対応する複数のキーワードが紐づいている辞書であり，これと前述の感情辞書を用いることで絵文字に対しても感情を解析することができる．

3.3 Word2Vec

感情分析のために，感情辞書の他に Word2Vec を活用する．Word2Vec は分散表現モデルであり，これを用いることで単語同士の類似度を計算することができる．そのため，ツイート内の単語と感情辞書内の単語の類似度を計算するために用いた．

Word2Vec には株式会社ホットリンクから提供されている，日本語大規模 SNS+Web コーパスから作成した hottoSNS-w2v[8][6] を用いた．この Word2Vec を構築したコーパスには SNS データが含まれているため本実験に適していると考えられる．

また，Word2Vec を利用するときにハイパーパラメータの設定を行う．詳細は表 3 に載せる．この 6*3=18 種類のハイパーパラメータの中から最も精度の高いハイパーパラメータを決定し，それによる感情分析結果を基に短期間イベントの分析を行う．精度の計算方法は，plutchik の 8 つの感情に対してそれぞれ，喜び・信頼・驚き・期待をポジティブとし，恐れ・悲しみ・嫌悪・怒りをネガティブとする．感情分析結果によりイベント発生時のそれぞれの感情ベクトル値を計算し，その比率がポジティブの方が高いほど精度が高いとする．これは目視により，イベント発生時の 9 割程度のツイートがネガティブツイートではないことが確認されているからである．

表 3 Word2Vec を適用するハイパーパラメータ

類似しているとみなす類似度閾値	0.5，0.6，0.7，0.8，0.9，1.0
Word2Vec を適用する品詞	[形容詞，副詞，動詞]，[形容詞，副詞]，[形容詞]

3.4 感情分析モデル

感情分析の全体の流れを以下に示す．

1. Plutchik の感情の輪に基づいて，ツイート毎に 8 つの感情に対応した 8 感情ベクトルを作成
2. ツイート内の単語と拡張した感情辞書を照合し，感情辞書に同単語が存在すれば，8 感情ベクトルに対し，その単語に紐づいた感情を 1 増加
3. 存在しなければ，絵文字キーワード対応辞書で絵文字をキーワードに変換し，そのキーワードに対して前述の拡張した感情辞書を実行
4. 存在しなければ，ハイパーパラメータに基づいて Word2Vec を活用し，拡張した感情辞書に存在する単語の中で最も類似度が高い単語を計算．その後 8 感情ベクトルに対し，その単語に紐づいた感情を 1 増加

また，感情辞書や Word2Vec で感情ベクトルの値を増加させるとき，否定語の対処を行った．日本語には語尾に「ない」が付くことで意味が反転する単語が多くあるため，本実験では Cabocha により作成した係り受けリストを用いて，「ない」が紐づいている単語，かつ感情辞書内に存在する単語の場合，感情ベクトルの値を 1 増加させる代わりに 1 減少させることで否定語の対処を行った．

結果としてツイート毎に 8 感情ベクトルを作成した．また Word2Vec のハイパーパラメータは，類似度閾値は 0.9，品詞は [形容詞，副詞] が最も精度が高かったため，これを用いた感情分析結果を考察に用いる．

[7] https://github.com/yagays/emoji-ja
[8] https://github.com/hottolink/hottoSNS-w2v

4　結果と考察

本章では 3 章の感情分析モデルを 2 章のデータセットに適用して作成した，感情ベクトルを分析することで考察を行う．

まず感情分析結果であるツイート毎の各感情ベクトルに対して Onehot 感情ベクトルを作成する．これは各感情ベクトルに対して，最も高い感情を 1，それ以外を 0 としたベクトルである．次に，全ツイートの感情ベクトルを時系列に 10 分毎もしくは 3 時間毎にまとめる．まとめ方は各感情毎の和を計算する．また 3 時間毎では 23 区毎に分割するものとそうでないものの 2 種類の感情ベクトルを作成する．結果として，10 分毎の 8 感情ベクトルと Onehot8 感情ベクトル，3 時間毎の 8 感情ベクトルと Onehot8 感情ベクトル，3 時間毎 23 区毎の 8 感情ベクトルと Onehot8 感情ベクトルの計 6 種類の感情ベクトルを作成した．この 6 つの感情ベクトルに対してそれぞれ分析を行う．

4.1　分析

分析方法として以下の 2 つを行う．

一つは感情の時系列の変化を見るため感情ベクトルから時系列 Z スコアを作成する．Z スコアとは，平均が 0，標準偏差が 1 になるように変換した値であり，時系列 Z スコアとは，ウィンドウサイズ毎に平均と標準偏差を計算した Z スコアを時系列に並べたリストである．これを用いることで各感情毎に時系列の変化がどのくらい発生しているかを分析することができる．

もう一つは同時刻の 8 つの感情を比較するため，感情ベクトルから感情割合ベクトルを作成する．このとき，8 つの感情の中で最も高い感情値を求め，それを分母として割合値を作成する．これを用いることで同時刻にどの感情が最も感情値が高いのかを把握でき，またその感情と他の感情との比較を行うことができる．

これらのグラフを短期間イベントの分析のために作成した．しかし，紙面の都合上全てのグラフを載せることができないため一部を載せた．それぞれ時系列 Z スコアは図 1，図 2，図 3 に，感情割合ベクトルは図 4，図 5，図 6 に示す．

4.2　考察

図 1 を見ると，ブルーインパルスイベントの発生時刻は 5/29/12:40-5/29/13:00 であるが，そのとき信頼はイベント直前から半日かけて上昇しているのに対し，喜びはイベント中のみ高くなっていたり，嫌悪や怒りはイベント前に大きく上昇したがイベント直前から大きく減少するなど，感情毎に異なる時間変化が見られる．よってブルーインパルスなどの短期間イベントにも，世論の感情への影響が存在すると考えられる．

また信頼以外の感情はイベント後は不安定だが，信頼の感情はイベント後半日かけて上昇している．図 4，図 5，図 6 の方を見ても，どれもイベント直前から信頼の感情が 8 つの感情の中で最も高くなっている．よってブルーインパルスは信頼の感情に対する影響が最も大きいと考えられる．

またイベント発生区域と非イベント発生区域に対する感情の影響に差異があるのか調べるためそれぞれ図 2，図 3，図 5，図 6 などで区域毎に分析をしたが，あまり差異は感じられなかった．そのため，区域毎に差異はない可能性があるがさらなる調査が必要である．

5　まとめ

本実験では 2020 年に東京で発生したブルーインパルスや日本中で発生した打ち上げ花火など，1 日未満の短期間イベントに対して，代表的な SNS であるツイッターを用いて感情分析することで SNS 上の世論の分析を行った．結果として，短期間イベントでも世論の感情に対して影響を与える可能性があることを示唆した．今後は区域毎の差異のさらなる調査，10 分や 3 時間だけでなく別の時間の長さでの分析，全く別のイベントに対して感情分析が適用可能かどうかの調査を行っていく．

参考文献

[1] Kirilenko, Andrei P., Svetlana O, et al: Sochi 2014 Olympics on twitter: perspectives of hosts and guests, Tourism Management, 63, pp. 54-65, 2017.

[2] Barnaghi, Peiman, Parsa Ghaffari, et al: Opinion mining and sentiment polarity on twitter and correlation between events and sentiment, 2016 IEEE second international conference on big data computing service and applications (BigDataService), pp. 52-57 2016.

[3] Jianqiang, Zhao, Xiaolin, et al: Comparison research on text pre-processing methods on twitter sentiment analysis, IEEE Access, 5, pp. 2870-2879, 2017.

[4] Zimbra, David and Abbasi, et al: The state-of-the-art in twitter sentiment analysis: a review and benchmark evaluation, ACM Transactions on Management Information Systems (TMIS), 9, 2, pp. 1-29, 2018.

[5] Plutchik Robert: The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice, American scientist, 89, 4, pp. 344-350, 2001.

[6] 松野省吾，水木栄，榊剛史: 日本語大規模 SNS+ Web コーパスによる単語分散表現のモデル構築，人工知能学会全国大会論文集 一般社団法人 人工知能学会，pp. 4Rin113-4Rin113，2019.

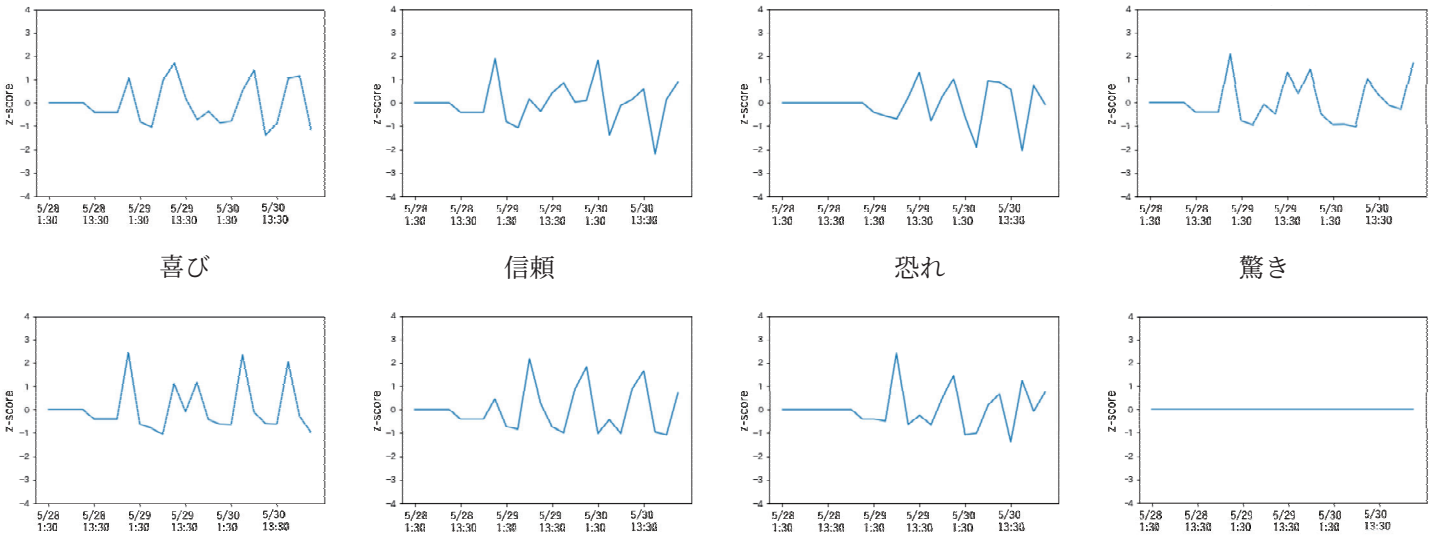


図 1 ブルーインパルスの 3 時間毎にまとめた onehot 感情ベクトルの時系列 Z スコアグラフ

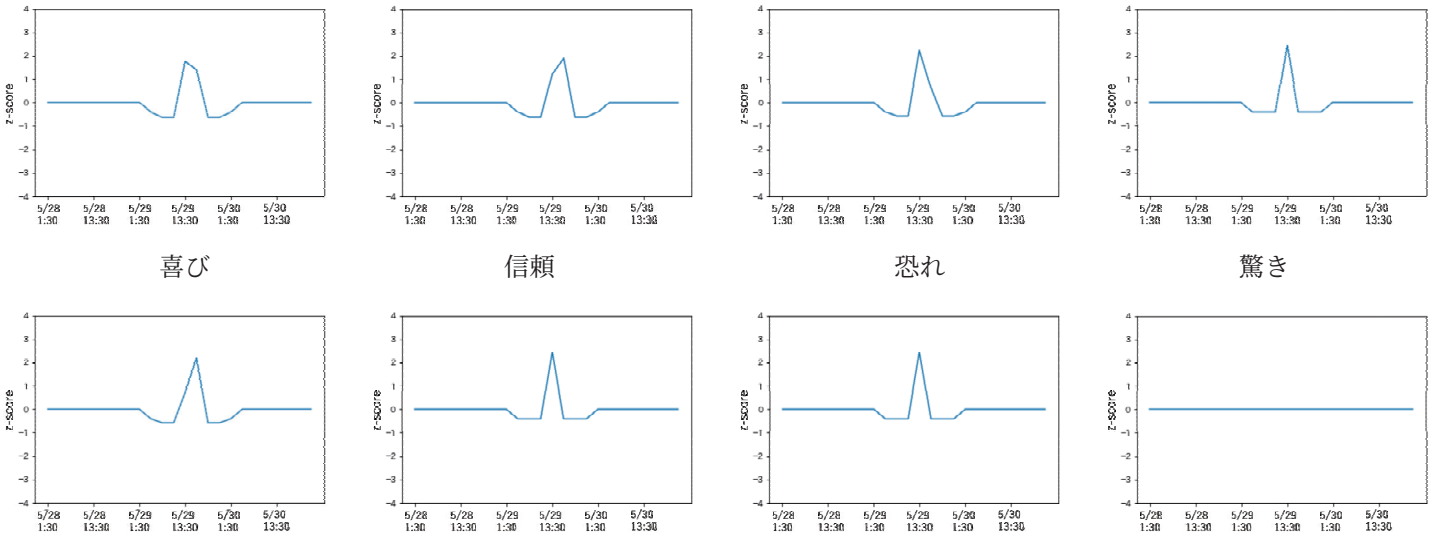
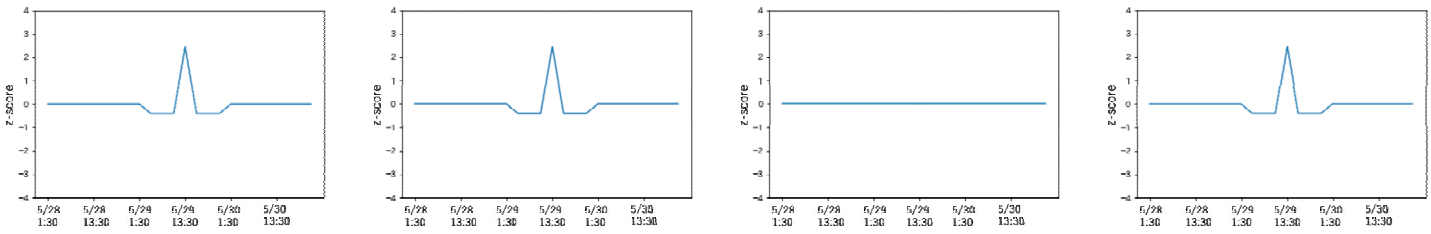
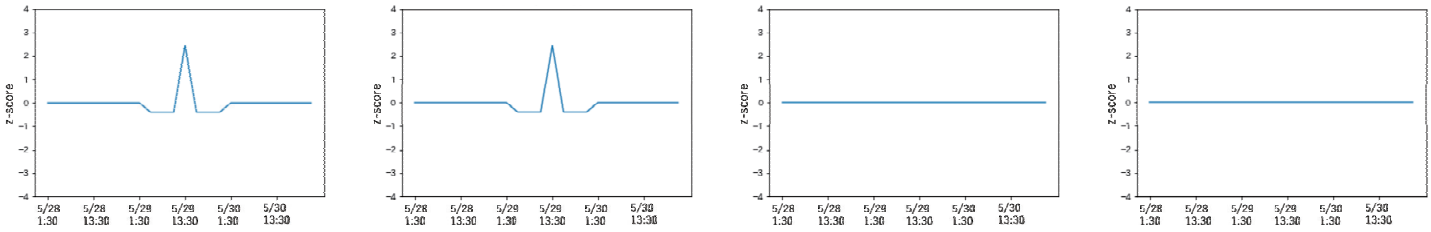


図 2　ブルーインパルスの 3 時間毎にまとめたイベント発生区域 (中央区) の onehot 感情ベクトルの時系列 Z スコアグラフ

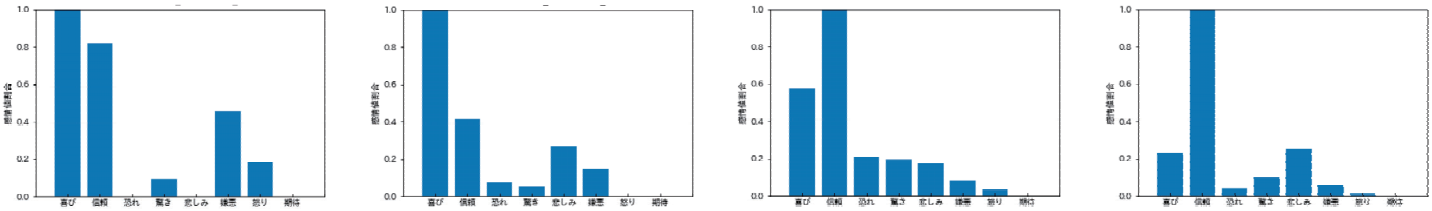


喜び　信頼　恐れ　驚き



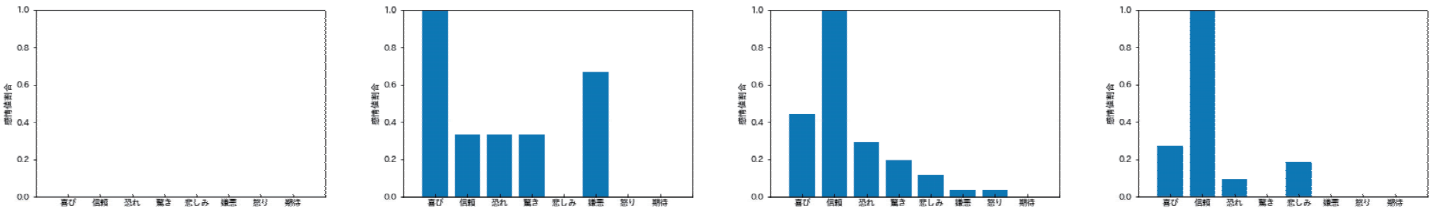
悲しみ　嫌悪　怒り　期待

図 3　ブルーインパルスの 3 時間毎にまとめた非イベント発生区域 (練馬区) の onehot 感情ベクトルの時系列 Z スコアグラフ



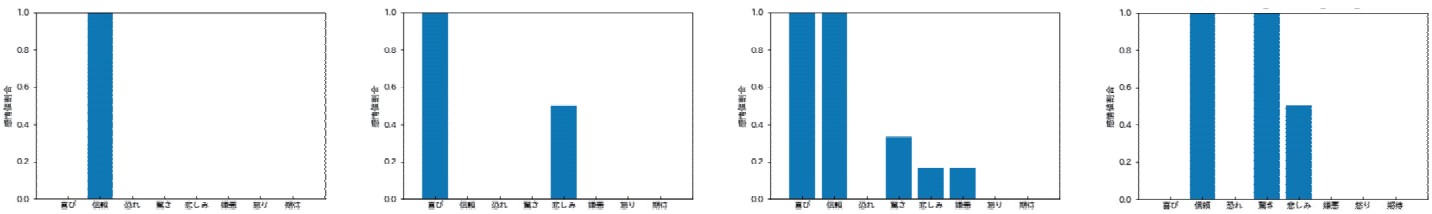
2020/5/29 6:00-9:00　2020/5/29 9:00-12:00　2020/5/29 12:00-15:00　2020/5/29 15:00-18:00

図 4 ブルーインパルスの 3 時間毎にまとめた onehot 感情ベクトルの感情割合グラフ



2020/5/29 6:00-9:00　2020/5/29 9:00-12:00　2020/5/29 12:00-15:00　2020/5/29 15:00-18:00

図 5 ブルーインパルスの 3 時間毎にまとめたイベント発生区域 (中央区) の onehot 感情ベクトルの感情割合グラフ



2020/5/29 6:00-9:00　2020/5/29 9:00-12:00　2020/5/29 12:00-15:00　2020/5/29 15:00-18:00

図 6 ブルーインパルスの 3 時間毎にまとめた非イベント発生区域 (練馬区) の onehot 感情ベクトルの感情割合グラフ

ARG WI2 No.16, 2020
WI2-2020-17

Instagram におけるアーカイブ投稿を引き起こす写真に関する基礎調査

中本 優希　比嘉 彗太　土方 嘉徳

関西学院大学商学部

contact@soc-research.org

概要　近年，Instagram は多くの人々によって利用されるようになった．Instagram では次々と新しい機能が追加されているが，2017 年 5 月にアーカイブ機能が追加された．アーカイブ機能とは，一度 Instagram 上に投稿及び公開した写真や動画を非公開にする機能のことである．この機能を用い，最近若者を中心に「アーカイブ投稿」という新しい投稿スタイルで写真や動画が投稿されるようになった．若者はこのアーカイブ投稿を，通常投稿（タイムラインに流れるように投稿するスタイル）に織り交ぜ，Instagram での投稿を楽しんでいる．我々は人に見られたい投稿スタイルである通常投稿と，投稿はしたいが人目につかないように投稿したいアーカイブ投稿では投稿される内容に違いがあるのではないかと考えた．本研究では，どのような写真がアーカイブ投稿を引き起こすのかを明らかにする．ペルソナを用いた被検者実験により，人が写る写真よりも人が写らない写真の方がアーカイブ投稿を引き起こしやすく，人が多く写る写真よりもより少ない人数で写る写真の方がアーカイブ投稿を引き起こしやすいことが分かった．

キーワード　Instagram，アーカイブ投稿，ユーザ行動分析

1 はじめに

近年，多くの人々によって Instagram が利用されている．Instagram では新しい機能が次々と追加されているが，2017年5月にアーカイブ機能が追加された．アーカイブ機能とは，一度 Instagram 上に投稿及び公開した写真や動画を非公開にする機能のことである．非公開にされた写真や動画は Instagram 上には保存されており，もう一度プロフィール上に戻すことができる．プロフィール上に戻す時，タイムラインの時系列は初めに投稿した時系列で公開される．

この機能を用い，最近若者を中心に「アーカイブ投稿」という新しい投稿スタイルで写真や動画が投稿されるようになった．「アーカイブ投稿」とは投稿した写真や動画をすぐにアーカイブ機能で非公開にし，数日後にもう一度プロフィール上に戻すことによってフォロワーが気づかないうちに写真や動画を投稿する投稿スタイルである．若者はこのアーカイブ投稿を，通常投稿（タイムラインに流れるように投稿するスタイル）に織り交ぜ，Instagram での投稿を楽しんでいる．

我々は，この研究を行うのに先立ち，事前調査として，どのくらいの人々がアーカイブ投稿を行っているのかを調査した．Instagram で便宜サンプリングにより抽出した 100 人のユーザを対象に，全投稿の中にアーカイブ投稿があるかどうかを調べた．ここで，アーカイブ投稿であるかどうかを厳密に判定することはできないが，通常投稿と思われる投稿の 2 割以下のいいね数しかついていないものをアーカイブ投稿とみなした．また，アーカイブ投稿を行っているユーザのうち，過去半年間の投稿の何割がアーカイブ投稿なのか調べた．調査の結果，アーカイブ投稿を行っているユーザは 32％であった．また，アーカイブ投稿を行っているユーザのうち，過去半年間でのアーカイブ投稿の割合の平均は 45％であった．このことから，アーカイブ投稿を行うユーザは少なからず存在することが分かる．

我々は人に見られたい投稿スタイルである通常投稿と，投稿はしたいが人目につかないように投稿したいアーカイブ投稿では投稿される内容に違いがあるのではないかと考えた．本研究では，どのような写真がアーカイブ投稿を引き起こすのかを明らかにする．

我々は，写真の特徴として以下の 5 つに注目した．Instagram のような写真を主体とした SNS では，投稿者が現実世界で経験したことを写真に撮り，投稿されることが多い．また，日常生活のありとあらゆる出来事を写真に撮るのではなく，より特別な出来事を写真に撮り，投稿する傾向がある．そのため，その写真に写っている「人」と，出来事の「特別感」に注目する．また，

Copyright is held by the author(s).
The article has been published without reviewing.

Instagram における投稿の主体は写真であるため，写真に何が写っているかも，人々の関心を惹くと思われる．そこで，写真の「撮影対象物」にも注目する．人々は，Instagram において現実世界の友人とつながっていることが多いと思われるが，投稿を誰に見てもらうのか，あるいは誰に見られるのかは重要な問題である．そこで，写真がある友人グループ内において撮影されたものかどうかを表す「コミュニティ」にも注目する．最後に，Instagram をファッションやアートなど，自分の好きな物やライフスタイルに基づき，他者からの共感を得るために使っている者もいる．そこで，撮影された写真に「趣味」の要素が入っているかどうかにも注目する．

上記の写真特徴に基づき，我々は以下の研究課題(Research Question)を立てる．

RQ1: 「人」の要素がアーカイブ投稿する・しないに関連があるのか
RQ2: 「コミュニティ」の要素がアーカイブ投稿する・しないに関連があるのか
RQ3: 「特別感」の要素がアーカイブ投稿する・しないに関連があるのか
RQ4: 「趣味」の要素がアーカイブ投稿する・しないに関連があるのか
RQ5: 「撮影対象物」アーカイブ投稿する・しないに関連があるのか

本論文の以降の構成は以下の通りである．2 章で関連研究を，3 章で実験方法を，4 章で実験析結果とそれに対する考察を述べる．最後に 5 章でまとめを述べる．

2 関連研究

オンライン上における印象形成に関する研究を 3 つ紹介する．

1 つ目は Seidman による Facebook における印象形成に関する研究である[1]．この研究は，Facebook を利用して帰属意識や自己表現のニーズを満たすことと，さまざまな自己の側面を表現することが，パーソナリティの一つであるビッグファイブとどのように関連しているかを調べている．この研究では，外向的で快活なユーザが Facebook をオフラインの関係を積極的に補完する方法として利用している可能性があることを見出し，また神経過敏な人は Facebook を隠れた理想的な自己表現ができる場として利用している可能性があることを明らかにした．

2 つ目は Salim らによる Instagram 上における印象形成に関する研究である[2]．この研究は，インスタグラムユーザの自己表現に対する友人関係に伴う自尊心と見逃すことの恐怖の影響を分析する研究である．自己表現に影響することが見出された唯一の変数は，見逃すことの恐怖のみであり，友人関係に伴う自尊心の影響は見られなかった．しかし，友人関係に伴う自尊心は見逃すことの恐怖には影響があることを明らかにした．

3 つ目は Brand らによる出会い系サイト上での印象形成に関する研究である[3]．この研究は，インターネット上の出会い系サイト内での男性のプロフィールにおける写真の魅力とテキストの魅力の相関を分析する研究である．この研究では，写真の魅力とテキストの魅力が相関していることが分かった．

また，Instagram 上における顔の有無とエンゲージメントに関する研究も 1 つ紹介する．現実世界におけるコミュニケーションにおいては，人間の顔が非言語コミュニケーションの重要なチャネルになっていることが明らかになっているが，Bakhshi らはオンライン上でも，このことが成り立つかを調査した[4]．顔の存在，年齢，性別が写真のソーシャルなエンゲージメントにどのような影響を与えるかを尋ね，二つの社会的エンゲージメントのフィードバック要因，「いいね!」とコメントを中心に調査した．その結果，顔のある写真は，「いいね!」を受け取る確率が 38%高く，コメントを受け取る確率が 32%高いことが分かった．しかし，顔の数，年齢および性別は影響しないことが分かった．

以上 4 つの関連研究より，人々はオンライン上においても人から良く見られようと，写真やテキストで自分をより理想的な形で表現する傾向があることが分かった．しかし，これらの研究はプロフィールや一般投稿に対して行われたものであるが，人目につかないように投稿可能なアーカイブ投稿を対象にしたものではない．本研究では，アーカイブ投稿を引き起こす写真の特徴について注目した初めての研究である．

3 実験方法

3.1 実験の概要

本研究ではアーカイブ投稿を引き起こす写真の特徴を被験者実験によって明らかにする．しかし，Instagram の利用頻度や利用方法は，性別や世代によって異なる可能性がある．そこで本研究では，実験対象者を女子大学生として，Instagram でアーカイブ投稿を行っている者に限定する． 実験内容としては，ペルソナのユーザを設定し，そのユーザを自分と自己同一視してもらうことで，そのペルソナユーザの撮影した写真と状況説明に対して，自分なら通常投稿するか，アーカイブ投稿するか，どちらもしないかを判断してもらった．

3.2 写真の特徴

本研究では，写真の特徴として，「人」，「コミュニティ」，「特別感」，「趣味」，「撮影対象物」の 5 つを対象とする．「人」は写真に人が写っているかどうかと，写真に写る人の数，「特別感」は写真に写る出来事がユーザにとって特別な出来事であるかどうか，「コミュニティ」は写真に

写る出来事が部活動やサークル，ゼミ(研究室)などコミュニティの中での出来事であるかどうか，「趣味」は趣味に関する写真であるかどうか，「撮影対象物」は写真の中に対象とするもの(今回の実験では食べ物，オブジェ，風景を用いる[5])が写っているかどうかを，特徴量とすることにした．なお，人の数は，0 人，1 人，2 人，3 人以上のカテゴリ値とした．

3.3 実験の詳細

Google Form を用いて実験を行った．図 1 のように実際に写真を撮るであろう場面を 24 個挙げた．実験では，被験者に，その場面で撮ったとする写真に対して，「アーカイブ投稿する」，「通常投稿する」，「どちらもしないがフォルダには残しておく」，「そもそもそのような行動はとらない」の 4 つの行動のうちの，どれを行うのかを回答してもらった．また，被検者に場面ごとのイメージを持ってもらいやすくするために，写真の例も添付した．

用意する場面は，「人」，「特別感」，「コミュニティ」，「趣味」，「撮影対象物」の 5 つの要素を組み合わせて決定した．ここで，「趣味」を「アイドルの追っかけ」とし，好きなアイドルのライブに行ったり，好きなアイドルの CD を購入するといった場面を想定した．「人」の人数，「特別感」の有無，「コミュニティ」の有無の 3 つに関しては表 1 に示されるラテン方格法を用い，組み合わせを決定した．「趣味」と「撮影対象物」の 2 つは，「コミュニティ」を含めてラテン方格法を用いた組み合わせをすると，現実的にあり得ない場面が出てきてしまった為，この 2 つに関しては，表 2，3 のように「人」の有無，「特別感」の有無のそれぞれとのすべての組み合わせによって組み合わせを決定した．例えば，趣味有り，人有り，特別感有りのような組み合わせである．

表 1 ラテン方格法(2 水準 2 因子 4 水準 1 因子)
人:1→0 人，2→1 人，3→2 人，4→3 人以上
特別感:1→あり，2→なし
コミュニティ:1→あり，2→なし

写真No.	人	特別感	コミュニティ
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1
5	3	1	1
6	3	2	2
7	4	1	2
8	4	2	1

表 2 趣味のカテゴリの組み合わせ

	趣味	人	特別感
9	あり	あり	あり
10	あり	なし	なし
11	あり	なし	あり
12	あり	あり	なし

表 3 撮影対象物のカテゴリの組み合わせ

	撮影対象物	人	特別感
13	食べ物	あり	あり
14	食べ物	なし	なし
15	食べ物	なし	あり
16	食べ物	あり	なし
17	オブジェ	あり	あり
18	オブジェ	なし	なし
19	オブジェ	なし	あり
20	オブジェ	あり	なし
21	風景	あり	あり
22	風景	なし	なし
23	風景	なし	あり
24	風景	あり	なし

(a) 画像例 1

(b) 画像例 2

図 1 実験で使用した画像の例

説明文の例を以下に 2 つ紹介する. また, これらの例で用いた写真を図 1 に示す.

(説明文例 1)
＜趣味あり, 人あり, 特別感あり＞
あなたは好きなアイドル(アーティスト)がいます。あなたはそのアイドル(アーティスト)のライブへ行きました。この写真はライブ会場前で一緒に行った友達とライブグッズを持って撮った写真です.

(説明文例 2)
＜人なし, 特別感あり, コミュニティあり＞
あなたはサークル(または部活, ゼミ)の合宿で金沢を訪れました。この写真は観光をする時間に撮った写真です.

3.4 分析方法

実験の分析は, カイ二乗分析を行ったのち, 残差分析を行い,「人」,「特別感」,「コミュニティ」,「趣味」,「撮影対象物」の 5 つそれぞれの要素がアーカイブ投稿をすること, 通常投稿すること, アーカイブ投稿・通常投稿のどちらもしないこと, の 3 つに関連があるのかどうかを調べた.

4 実験結果

得られたデータは, 重複したデータ(一人のユーザが 2 回以上回答しているデータ)を除いたところ, 分析対象となるデータは 96 であった.「人」の分析は人数(0 人, 1 人, 2 人, 3 人以上)の場合と, 人の有無の場合の 2 パターンで分析を行った. その他 4 つの要素は, 全て有無で分析を行った.「撮影対象物」に関しては「撮影対象物」である「食べ物」,「オブジェ」,「風景」の 3 つのいずれかが写っている場合と写っていない場合で分けたものと,「食べ物」,「オブジェ」,「風景」のそれぞれで対象物が写っている場合と写っていない場合で分けたもので分析を行った.

4.1「人」の人数での比較

「人」の人数において, カイ二乗分析を行ったところ, χ^2=64.222, p値=6.218e-12＜0.05, df=6 となり, 有意差が見られた. さらに残差分析を行ったところ,「人 0 人・アーカイブ投稿」と「人 1 人・アーカイブ投稿」において, 調整済み残差が 2.08, 2.55 > 1.96 となり, 有意差が確認できた.「人 3 人以上・通常投稿」においても調整済み残差が 6.16 > 1.96 となり, 有意差が見られた. これらのことから, 人が 0 人か 1 人写る写真はアーカイブ投稿になりやすく, 人が 3 人以上写る写真は通常投稿になりやすくなることが確かめられた.

4.2「人」の有無での比較

「人」の有無において, カイ二乗分析を行ったところ, χ^2=7.3087, p値=0.02588<0.05, df=2 となり, 有意差が見られた. さらに残差分析を行ったところ,「人なし・アーカイブ投稿」において, 調整済み残差が 2.08 > 1.96 となり, 有意差が確認できた. また,「人あり・通常投稿」においても調整済み残差が 2.15 > 1.96 となり, 有意差が見られた. これらのことから, 人が写らない写真はアーカイブ投稿になりやすく, 人が写る写真は通常投稿になりやすくなることが確かめられた.

4.3「特別感」の有無での比較

「特別感」の有無において, カイ二乗分析を行ったところ, χ^2=128.34, p値=2.2e-16<0.05, df=2 となり, 有意差が見られた. さらに残差分析を行ったところ,「特別感あり・通常投稿」において, 調整済み残差が 10.156 > 1.96 となり, 有意差が見られた. また,「特別感なし・どちらもしない」においても調整済み残差が 10.826 > 1.96 となり, 有意差が見られた. これらのことから, 特別感がある写真は通常投稿されやすく, 特別感がない写真はそもそも投稿されにくいことが確かめられた.

4.4「コミュニティ」の有無での比較

「コミュニティ」の有無において, カイ二乗分析を行ったところ, χ^2=45.985, p値=1.034e-10<0.05, df=2 となり, 有意差が見られた. さらに残差分析を行ったところ,「コミュニティあり・アーカイブ投稿」において, 調整済み残差が 3.11 > 1.96 となり, 有意差が見られた. このことから, コミュニティに関連する写真はアーカイブ投稿になりやすいことが確かめられた.

4.5「趣味」の有無での比較

「趣味」の有無において, カイ二乗分析を行ったところ, χ^2=11.041, p値=0.004004<0.05, df=2 となり, 有意差が見られた. さらに残差分析を行ったところ,「趣味なし・通常投稿」において, 調整済み残差が 2.45 > 1.96 となり, 有意差が見られた. また,「趣味あり・どちらもしない」においても調整済み残差が 3.21 > 1.96 となり, 有意差が見られた. これらのことから, 趣味が見られない写真は通常投稿になりやすく, 趣味に関連する写真はそもそも投稿されにくいことが確かめられた.

4.6「撮影対象物」の有無での比較

「撮影対象物」の有無において, カイ二乗分析を行ったところ, χ^2=11.041, p値=0.004004<0.05, df=2 となり, 有意差が見られた. さらに残差分析を行ったところ,「撮影対象物あり・通常投稿」において, 調整済み残差が 2.45 > 1.96 となり, 有意差が見られた. また,「撮影対象物なし・どちらもしない」においても調整済み残差が 3.21 > 1.96 となり, 有意差が見られた. これらのことから, 撮影対象物が写る写真は通常投稿になりやすく, 撮影対

象物が写らない写真はそもそも投稿されにくいことが確かめられた.

4.7 「撮影対象物」の種類での比較

「撮影対象物」の「食べ物」,「オブジェ」,「風景」の 3 つそれぞれにおいて, カイ二乗分析を行ったところ, χ^2 =14.708, p値=0.005347<0.05, df=4 となり, 有意差が見られた. さらに残差分析を行ったところ,「オブジェ・通常投稿」において, 調整済み残差が 2.41 > 1.96 となり, 有意差が見られた. また,「食べ物・どちらもしない」においても調整済み残差が 3.25 > 1.96 となり, 有意差が見られた. これらのことから, オブジェが写る写真は通常投稿されやすく, 食べ物が写る写真はそもそも投稿されにくいことが確かめられた.

4.8 考察

実験結果から, 仮説に対しての考察を行った.

まず, 仮説 1 に対して, 人が写る写真よりも人が写らない写真の方がアーカイブ投稿を引き起こしやすく, 人が多く写る写真よりもより少ない人数で写る写真の方がアーカイブ投稿を引き起こしやすいことが分かった. また, 特に人が 1 人で写る場合が最もアーカイブ投稿を引き起こしやすい. これは, 人が 1 人で写っている場合というのは投稿者本人が写っている可能性が高いと考え, 恥ずかしさからアーカイブ投稿を引き起こすのではないかと考える.

仮設 2 に対して, 特別感がある写真は投稿自体は引き起こしやすいが, アーカイブ投稿よりも通常投稿を引き起こしやすいことが分かった. これは, 特別な出来事を他人の目につく形でアピールしたいという気持ちから通常投稿を引き起こすのではないかと考える.

仮説 3 に対して, コミュニティが見られる写真がアーカイブ投稿を引き起こしやすいことが分かった. コミュニティの人は Instagram での繋がりがあり, 共に同じ出来事を過ごした場合, タイムラインに同じような投稿が溢れてしまいフォロワーに飽きられるのではないかと思うためにアーカイブ投稿になりやすいのではと考える.

仮設 4 に対して, 趣味がある写真はそもそも投稿自体を引き起こしにくく, 趣味がない写真は通常投稿を引き起こしやすいことがわかった. 趣味が分かる写真は, 他人から共感してもらえない可能性がある為, 投稿をしづらいのだと考える.

仮設 5 に対して, 撮影対象物が写る写真は通常投稿を引き起こしやすいことが分かった. また, オブジェが写る写真は通常投稿されやすく, 食べ物が写る写真はそもそも投稿されにくいことが分かった. 食べ物を撮影するのは Instagram に投稿する為に撮影するというよりも, 記録としてカメラフォルダに残しておくために撮影すると考える為, 撮影対象物の中でも食べ物はそもそも投稿されにくいのだと考える.

4.9 実験の制約

今回の実験での写真の添付は, あくまで参考程度のものと考え我々は写真を実験の際に添付したが, 被験者にとっては写真の影響は大きく, 写真のクオリティによって実験結果に影響が出た恐れがある. また, 想定される場面のパターンが少なく, 被験者にとって経験しないような場面しか提示できなかったことによって,「そもそもそのような行動はとらない」の回答に偏ってしまう場面もいくつか見られた. 例えば, 趣味があり人が写る特別感のない写真は,「そもそもそのような行動はとらない」の回答が多かった. 今回用意した写真は「購入した CD を持って写る自分の写真」であったが, 購入した CD のみの写真は撮影したとしても, 自分が持っているところを自撮りするという行動は一般ユーザはしない行動であったのかもしれない. このことから, 特に「趣味」に関しては, 人によって経験しているものが様々であり, いくつかのパターンを用意すべきであった.「趣味」では, 今回「アイドルの追いかけ」という趣味を提示し, その場面を自分の趣味に置き換えて考えてもらうように提示していたが, 被検者の持つ趣味と「アイドルの追いかけ」という趣味があまりにもかけ離れた趣味ならば, 被験者にとってそれは自分の趣味に置き換えるということは難しかったかもしれない. 今後は,「趣味」に関して被検者の大多数が持つ「趣味」に少しでも近い「趣味」が提示できるように想定される場面のパターンを増やし, 写真のクオリティも統一されたものを用意して再度実験を行っていきたい.

5 まとめ

本研究では, ペルソナ実験においてアーカイブ投稿を引き起こす写真の特徴について明らかにした. 具体的には, 写真に写る人の人数が少ない, または写る人がいない方がアーカイブ投稿引き起こしやすく, 特に 1 人しか写らない場合が最もアーカイブ投稿を引き起こしやすいことが分かった. また, コミュニティが見られる写真もアーカイブ投稿を引き起こしやすいことが分かった.

今後は, 想定される場面のパターンを増やし, 写真のクオリティも統一されたものを用意して再度実験を行っていきたい. また, 写真の特徴だけではなく, パーソナリティもアーカイブ投稿を引き起こす要因となりうるのかを明らかにしていたいと考えている.

6 参考文献

[1] Seidman, G.: Self-presentation and Belonging on Facebook: How Personality Influences Social Media Use and Motivations, Personality and Individual Differences, Vol. 54, pp. 402-407, 2013.

[2] Salim, F., Rahardjo, W., Tanaya, T., et al.: Are Self-Presentation of Instagram Users Influenced by

Friendship-Contingent Self-Esteem and Fear of Missing Out?, Makara Hubs-Asia, Vol. 21, No. 2, pp. 70-82, 2017.

[3] Brand, R. J., Bonatsos, A., D'Orazio, R., et al.: What is Beautiful is Good, Even Online: Correlations between Photo Attractiveness and Text Attractiveness in Men's Online Dating Profiles, Computers in Human Behavior, Vol. 28, pp. 166-170, 2012.

[4] Bakhshi, S., Shamma, D. A. and Gilbert, E.: Faces Engage Us: Photos with Faces Attract More Likes and Comments on Instagram, Proc. of CHI'14, pp. 965-974, 2014.

[5] Kayako Morimoto and Yoshinori Hijikata: The Relationship between Photograph Subject and "Like!" Acquisition in Instagram, Proc. of IPSJ the 26th International Conference on Collaboration Technologies and Social Computing (CollabTech'20), Poster Proceedings, 2020.

ARG WI2 No.16, 2020
WI2-2020-18

バーストを用いた論文の特性分析

小林 和央[†,a]　風間 一洋[†,b]　吉田 光男[††,c]　大向 一輝[†††,d]
佐藤 翔[††††,e]　桂井 麻里衣[†††††,f]

† 和歌山大学大学院システム工学研究科　†† 豊橋技術科学大学情報・知能工学系
††† 東京大学大学院人文社会系研究科　†††† 同志社大学 免許資格課程センター　††††† 同志社大学理工学部
a) *s206099@wakayama-u.ac.jp*　b) *kazama@wakayama-u.ac.jp*　c) *yoshida@cs.tut.ac.jp*　d) *i2k@l.u-tokyo.ac.jp*
e) *min2fly@slis.doshisha.ac.jp*　f) *katsurai@mm.doshisha.ac.jp*

概要　ソーシャルメディア上の反応を用いるオルトメトリクスは，既存の被引用数などの論文評価指標との相関が低く，異なる種類のユーザ評価が混在していると考えられる．本稿では，ユーザに応じた論文評価指標の実現のために，インターネット上の情報拡散や学術活動由来のバーストと，オルトメトリクスに用いられる評価指標との関連を分析する．具体的には，ソーシャルメディア上の論文言及数と論文検索サービス上の論文閲覧数に関する行動的バーストと，論文検索システムの検索語に関する意味的バーストを抽出し，バーストの有無や数，期間と各指標との関連性を分析する．実際に，2 年間のソーシャルメディア上の論文言及データと CiNii Articles の検索ログを用いて，バースト論文は各指標が高くなりやすい傾向や，言及バーストは閲覧バーストを引き起こしやすいが言及由来ではないバーストも存在すること，意味的バーストする単語の傾向の違いなどを明らかにする．

キーワード　CiNii Articles, ソーシャルメディア, オルトメトリクス, バースト, 2 部グラフ

1　はじめに

計量書誌学の分野では，今まで論文の評価指標として被引用数が主に用いられてきた．ただし，引用数には評価遅延があり，一般的に論文公開から引用数が増えるまで 2〜3 年掛かると言われている [1]．そこで，文献の閲覧数，ソーシャルメディアの言及，マスメディアの報道など，研究者以外の関与を含めた社会的な影響を示す様々な視点を組み入れることで，文献が社会に及ぼした影響度を包括的かつ早期に計測することを目指す指標であるオルトメトリクス（Altmetrics）が提案されている．ただし，単純に被引用数の代わりにオルトメトリクスを使えるわけではなく，Priem らはオルトメトリクスと被引用数の相関が弱いことを指摘し，オルトメトリクスは被引用数とは異なるユーザの影響を反映していると述べている [2]．すなわち，被引用数は論文を執筆する研究者に限定された評価であるのに対し，オルトメトリクスは広範・種々のユーザによる評価に基づくと考えられることから，例えば，異なる種類のユーザ評価を分離できれば，被引用数と相関が高いオルトメトリクスを実現したり，ユーザの種別に応じて論文評価指標を用いることが可能になると考えられる．

本稿では，インターネット上の情報拡散や学会などの学術活動，教育機関によって引き起こされるバースト現象が，オルトメトリクスなどの評価指標に与える影響を分析する．具体的には，まずソーシャルメディアの論文言及人数と論文検索サービスである CiNii Articles 上の論文閲覧人数を論文の評価指標として，その時系列変化から行動的バーストを抽出する．さらに，各論文の意味的な推定のために，CiNii Articles の利用履歴から検索語と検索人数を抽出し，その時系列変化から意味的バーストを抽出する．論文の評価指標とバーストの有無や数，期間の関連性を分析し，ユーザに応じた論文評価指標の実現に利用できるかを検討する．

2　関連研究

2.1　オルトメトリクスに関する分析

オルトメトリクスの特性に関する分析として，Priem らは，オルトメトリクスと被引用数の相関が弱く，オルトメトリクスは被引用数とは異なるユーザのインパクトを反映していると述べている [2]．佐藤らは，日本の学協会が発行する学術雑誌に掲載された論文を対象に，ソーシャルメディア上での言及数の分布や，論文の出版年，記述言語や属する分野等と言及数の関係を分析した [3]．日本の学協会誌掲載論文約 110 万件に対して，ソーシャルメディア上での言及が存在する論文は約 1%と少なかった．さらに，論文の言及数に分野による有意差が部分的に存在し，Ceek.jp Altmetrics と Altmetric では，どちらも人文社会系の論文が多く，理工系の論文が少ないという有意差があった．佐藤らは，Twitter からの言及数が多い論文と 0 回の論文集合について，非専門家にとっての論文タイトルの面白さを 7 段階で得点化し，言及数との相関を分析した [4]．その結果，得点と言及数には有意な相関は見られなかったが，言及数が多い論文集合のほうが 0 回の集合よりも得点が有意に高かったと述べ

Copyright is held by the author(s).
The article has been published without reviewing.

ている．

また，ソーシャルメディアのデータ利用に関する問題点に関する研究として，Mohammadi らは，学術目的で Twitter を利用しているユーザ 1,912 人にアンケート調査を行った結果，そのうち 45%は非学術系の職業のユーザで，59%が社会科学や人文学系のユーザであった [5]．さらに，Twitter の利用理由を 1,811 人に調査した結果，情報共有目的には 66%が，学術イベント等でのコミュニケーション目的には 52%が同意していたが，科学出版物の結果を公に伝える目的のユーザは 47%で，教育目的で利用しているユーザは 16%と少なかった．

つまり，被引用数と異なり，ソーシャルメディアに基づく指標は，一般ユーザ向けの話題に反応が偏る可能性や，論文の質ではなく社会に与えた影響を表すことに注意する必要がある．

2.2 バースト検知に関する研究

Kleinberg は，特定のイベントに関する文書の発生頻度の急激な増加現象であるバーストを検知する連続型と列挙型の 2 種類のアルゴリズムを提案した [6]．連続型では，時間軸に沿って断続的に発生する関連文書の時刻を元に，時間間隔がそれ以前と比べて短い状態が続くとバーストと判定する．列挙型では，時間軸を適当な間隔（バッチ）に区切り，バッチに含まれる関連文書数が全文書数と比べて多い状態が続くとバーストと判定する．連続型は観測粒度は細かくても同時刻の複数の関連文書の発生は扱えないが，列挙型は観測粒度は粗くても同時刻の複数の関連文書の発生を扱える特徴を持つ．

バースト検知をソーシャルメディア分析に用いた関連研究として，佐藤らは，テレビドラマに関する話題の盛り上がりを，4 種類のドラマを対象に，Twitter での反応のバーストに結びつくツイート内容から，頻出する特徴語を抽出して分析した [7]．その結果，各ドラマごとのバーストについて，それぞれ特定の特徴語と関連し，異なる影響でバーストすると述べている．佐々木らは，Twitter でユーザが注目したイベントに関する単語を，バーストの出現周期に着目して分類した [8]．その結果，異なる周期でバーストする単語では，関係する時間帯や毎週の番組など，単語の特徴が異なることを示した．

つまり，バーストに関連する特徴語やバーストの期間や周期性に着目することで，論文の特性の違いを分析できると考えられる．

3 バーストを用いた論文の特性分析手法

3.1 評価指標とバースト

時系列表現可能な評価指標を分析対象とする．時系列表現可能とは，指標 M が離散時間 $t(1 \leq t \leq N)$ における指標値 m_t の総和として $M = \sum_{t=1}^{N} m_t$ で求められる場合であり，その元となる $(m_0, m_1, \ldots, m_N)$ を M の時系列ベクトル表現と呼ぶ．

このベクトルには最終的な評価指標には現れない，指標とは直交する特性が埋め込まれていると仮定する．特性としては，例えば季節性，周期性，トレンド，同期性などが考えられるが，特にバーストに注目する．この理由は，例えば同じ指標値であったとしても，それが研究内容の評価が時間の経過と共に少しづつ蓄積されたからか，それともソーシャルメディア上でバズったことによる短期間のバーストかによって，その値の意味が大きく異なると考えられるからである．

3.2 行動的バーストと意味的バースト

論文に対する行動から求める重要性の評価指標に関するバーストである**行動的バースト**と，論文の内容を表すキーワードの利用数に関するバーストである**意味的バースト**の 2 種類を分析に用いる．

行動的バーストの分析には，ソーシャルメディア上の論文に関する言及人数と，CiNii Articles 上の論文の書誌情報の閲覧人数を用いる．言及人数の場合は，例えばソーシャルメディア上で論文に関する話題がバズった場合に，論文の URL や題名を含む言及ツイートのバーストが生じると考えられる．閲覧人数の場合は，例えば実世界で研究や教育に関連するイベントが生じた場合に，CiNii Articles や Google などの論文の検索が増大し，論文の書誌情報の閲覧のバーストが生じると考えられる．ただし，言及ツイートの URL をクリックしても論文を閲覧できるために，閲覧人数は言及人数の影響を少なからず受けると推測される．

意味的バーストの分析には，CiNii Articles や Google で論文を検索して，実際に書誌情報を閲覧した時に用いていた検索語の使用人数を用いる．ただし，言及ツイートの URL をクリックしただけの場合は考慮されないことに注意する．この場合には，実世界の何らかのイベントやマスメディアの報道をきっかけとして，一時的に論文の著者や内容を指定した検索が増加することでバーストが生じる．なお，本稿では論文の書誌情報が閲覧された理由を推測するために検索語を使用するが，意味的バーストは論文ではなく，論文を探すために使われた検索語が定常的か一時的かを判断するために用いる．

3.3 列挙型バースト検知アルゴリズム

例えば時系列表現可能な評価指標を対象にすること，Twitter では同時刻のツイートの存在可能性が高いこと，時系列データによっては正確な文書発生時刻がわからないことなどを考慮して，Kleinberg の列挙型バースト検知アルゴリズムを用いる [6]．

離散時間 $t(1 \leq t \leq N)$ に文書集合 $A_1, A_2, \ldots, A_N$ が送られてくるとして，対象単語 w を含む関連文書と含

まない非関連文書に分類する．時間 t の文書集合 A_t の文書数を d_t，対象単語 w を含む関連文書数を r_t とすると，解析期間全体の全文書数は $D = \sum_{t=1}^{N} d_t$，単語 w を含む全文書数は $R(w) = \sum_{t=1}^{N} r_t$ となる．

列挙型バースト検知アルゴリズムは，q_0 を非バースト状態，q_1 をバースト状態とする 2 状態オートマトンをモデルとして用いる．非バースト状態 q_0 に解析期間全体の単語 w の出現確率 $p_0 = R/D$，バースト状態 q_1 に p_0 にパラメータ s を掛けた値 $p_1 = sp_0$ とする．ただし，s は，$s > 1$ かつ $p_1 \leq 1$ を満たす値とする．なお，この s の値が小さいほど，文書集合中の関連文書の割合が低くてもバーストとみなされやすくなる．

状態遷移は d_t と r_t を入力として決定され，状態系列 $\mathbf{q} = (q_{i_1}, \ldots, q_{i_N})$ と表す．q_{i_t} は時間 t の文書集合によって決定された状態 $q_i (i = 0, 1)$ である．2 状態オートマトンにおいて，非バースト状態の q_0 とバースト状態の q_1 であることに対するコストを計算する．対象単語 w を含む文書が 2 項分布 $B(d_t, p_i)$ に従って出現すると仮定すると，状態 q_i であることに対するコスト関数 $\sigma(i, r_t, d_t)$ は次式で定義できる．

$$\begin{aligned} \sigma(i, r_t, d_t) &= -\ln [B(d_t, p_i)] \\ &= -\ln [\begin{pmatrix} d_t \\ r_t \end{pmatrix} p_i^{r_t} (1 - p_i)^{d_t - r_t}] \quad (1) \end{aligned}$$

ここで，$\begin{pmatrix} d_t \\ r_t \end{pmatrix} = C(d_t, r_t) = \dfrac{d_t!}{r_t!(d_t - r_t)!}$ である．つまり，時間 t の文書数 d_t から対象単語 w を含む関連文書数 r_t を選ぶ組み合わせである 2 項係数を表す．この関数は，入力 r_t と d_t，$q_i (i = 0, 1)$ によってコストが決まる．$p_1 > p_0$ であり，時間 t の関連文書の出現確率 r_t/d_t が p_1 に近ければ $\sigma(1, r_t, d_t) < \sigma(0, r_t, d_t)$ となりバースト状態 q_1 が，逆に r_t/d_t が p_0 に近ければ $\sigma(1, r_t, d_t) > \sigma(0, r_t, d_t)$ となり非バースト状態 q_0 が選ばれる．

ただし，バースト状態と非バースト状態が頻繁に切り替わらないように，状態 q_i から q_j への状態遷移を妨げる関数 $\tau(i, j)$ を次のように定義する．

$$\tau(i, j) = \begin{cases} (j - i)\gamma & (j > i) \\ 0 & (j \leq i) \end{cases} \quad (2)$$

τ はパラメータ γ によって調整されるが，特に理由がなければ，デフォルト値である $\gamma = 1$ とする．

最後に，状態 $q_i (i = 0, 1)$ であることに対するコスト関数 $\sigma(i_t, r_t, d_t)$ と，頻繁にバースト状態と非バースト状態の状態遷移が起こらないようにする $\tau(i_t, j_{t+1})$ より，次のコスト関数を最小とする状態系列 $\mathbf{q}$ を求めて，対象単語 w のバースト状態を表す状態系列とする．

$$c(\mathbf{q}|r_t, d_t) = \sum_{t=0}^{N-1} \tau(i_t, i_{t+1}) + \sum_{t=1}^{N} \sigma(i_t, r_t, d_t) \quad (3)$$

本稿では，ソーシャルメディア上の論文の言及人数と CiNii Articles 上の論文の書誌情報の閲覧人数，書誌情報を閲覧するために用いた検索語の使用人数の時系列表現に対して列挙型バースト検知アルゴリズムを適用し，得られたバースト期間リストを分析に使用する．

4 データセット

4.1 ソーシャルメディアの論文の言及データ

収集対象とする論文アーカイブの URL の一部をクエリとして，対象のソーシャルメディアサービスで提供される検索 API を利用し，個々の論文の URL とそれに対する言及テキストを取得した [9]．例えば，CiNii Articles に収録されている論文は `http://ci.nii.ac.jp/naid/110008898261` のような URL を持つので，すべての論文に共通する `ci.nii.ac.jp` をクエリとする．また，ソーシャルメディアサービスとして，Facebook，Google+，Twitter，OKWave，Yahoo!知恵袋，CiteULike，Delicious，はてなブックマーク，Wikipedia，レファレンス協同データベースの計 10 サイトを対象とした．

ユーザの識別子として，各ソーシャルメディアサービス名とサービスで使われているユーザ名の組を用いた．このために，同じユーザ名であっても，異なるサービスでは別のユーザとして扱われる．ただし，bot ユーザは分析に大きな影響を与えるために，Twitter の言及数が上位の投稿を人手で確認し，bot と判定した場合は除外した．論文の識別子として，論文の NAID（NII 論文 ID）を用いた．言及データには CiNii Books の書籍に割り当てられている NACSIS-CAT 書誌 ID である NCID（NII 書誌 ID）も含まれるが，NAID を含む言及データのみを使用した．対象期間は 2014 年 4 月 1 日〜2016 年 3 月 31 日の 2 年間で，最終的な言及論文の総数は 31,157 件，総ユーザ数は 29,685 人であった．

4.2 CiNii Articles の論文の閲覧データ

Apache Web サーバのアクセスログから，佐藤の手法 [10] を参考にクローラ等の機械的アクセスを除去してから，ユーザと閲覧論文のデータを抽出した．

ユーザの識別子として，IP アドレスとユーザエージェントの組 [11] を用いた．一般的に HTTP Cookie を利用してユーザ識別子を割り振ることが多いが，CiNii のアクセスログには Cookie が記録されていないためである．ただし，この方法には DHCP や NAT により別のユーザにも同じ IP アドレスが割り当てられている可能性や，Web ブラウザの更新でユーザエージェントが変わること

で，同一ユーザが別のユーザとして認識される可能性があることに注意する．論文の識別子として，URL に含まれる NAID（NII 論文 ID）を用いた．対象期間は 2014 年 4 月 1 日～2016 年 3 月 31 日の 2 年間で，最終的な閲覧論文の総数は 9,106,860 件，ユーザ数は 13,038,381 人であった．

4.3 論文の検索語データ

CiNii Articles のアクセスログから，論文のトピックを示す検索語を以下の手順で抽出する．

1. 閲覧論文のリファラが CiNii Articles や Google の検索エンジンであった場合に，その URL からクエリを抽出する．
2. 英大文字を英小文字に変換し，mecab-ipadic-NEologd の方式[1]で正規化する．
3. MeCab で形態素解析し 2 文字以上の名詞を抽出する．既存の IPA 辞書では認識できない固有表現に対応するため，mecab-ipadic-NEologd を利用する．
4. Oracle Text で提供されるストップワードリスト[2]に含まれる英単語を除去する．

論文 $p_i (i = 1, \ldots, N)$ の検索語は，論文 p_i の検索語 t_j の使用頻度 $qf(i,j)$ と使用文書数 $df(t_j)$ から，(4) のように $qf\text{-}idf(i,j)$ で重み付けする．

$$qf\text{-}idf(i,j) = \frac{qf(i,j)}{s(p_i)} \log \frac{N}{df(t_j)} \tag{4}$$

ここで $s(p_i)$ は文書 p_i の全検索語の使用頻度の和である．

各論文の上位の検索語を，その論文の内容の特徴を的確に表すキーワードとして使用する．

4.4 時系列データからのバースト検知

論文の言及・閲覧データから，論文の NAID を照合して両方に含まれる論文 30,679 件を抽出して，2014 年 4 月 1 日～2016 年 3 月 31 日の 2 年間の論文の言及人数と閲覧人数，バースト期間リストを抽出した．なお，論文の時系列ベクトルは 7 日単位で作成した．この理由は，1 日単位だと疎すぎることと，人間の基本的な生活周期である 1 週間に合わせることで，平日・休日の差などの影響を除去するためである．さらに，検索語に対してもバースト期間リストを抽出し，1 回でもバーストした場合にバースト検索語とした．

なお，バースト検知のパラメータは，通常はバーストのなりにくさを表すパラメータを $s = 2.0$，バースト状態への移行コストを $\gamma = 1$ とすることが多いが，論文ごとの閲覧人数や言及人数が少なくても，全体の頻度が疎な時期に，バースト判定されるのを防ぐために，$s = 4.0$，$\gamma = 1$ とし，全ての時系列データで同じ設定を用いた．

[1] https://github.com/neologd/mecab-ipadic-neologd/wiki/Regexp.ja

[2] https://docs.oracle.com/cd/E16338_01/text.112/b61357/astopsup.htm#i634475

表 1 閲覧・言及バースト論文のクロス集計結果

	言及バースト（有）	言及バースト（無）	合計
閲覧バースト（有）	2,174	7,437	9,611
閲覧バースト（無）	621	20,447	21,068
合計	2,795	27,884	30,679

5 論文の特性分析

5.1 バースト論文の関係分析

まず，言及・閲覧バースト論文の関係について分析する．各バースト論文数をクロス集計した結果を，表 1 に示す．言及バースト論文数は 2,795 件で，そのうち 2,174 件（77.9%）が閲覧でもバーストしていた．この結果から，ソーシャルメディア上の言及をきっかけに，同時に多くの閲覧が引き起こされていることが確認できた．

また，閲覧バースト論文数は 9,611 件であり，言及バースト論文数の 3.4 倍であった．この結果から，ソーシャルメディア上の情報拡散由来ではなく，研究や学習関連のイベント由来のバーストが存在すると推測できる．

5.2 バースト論文の順位傾向分析

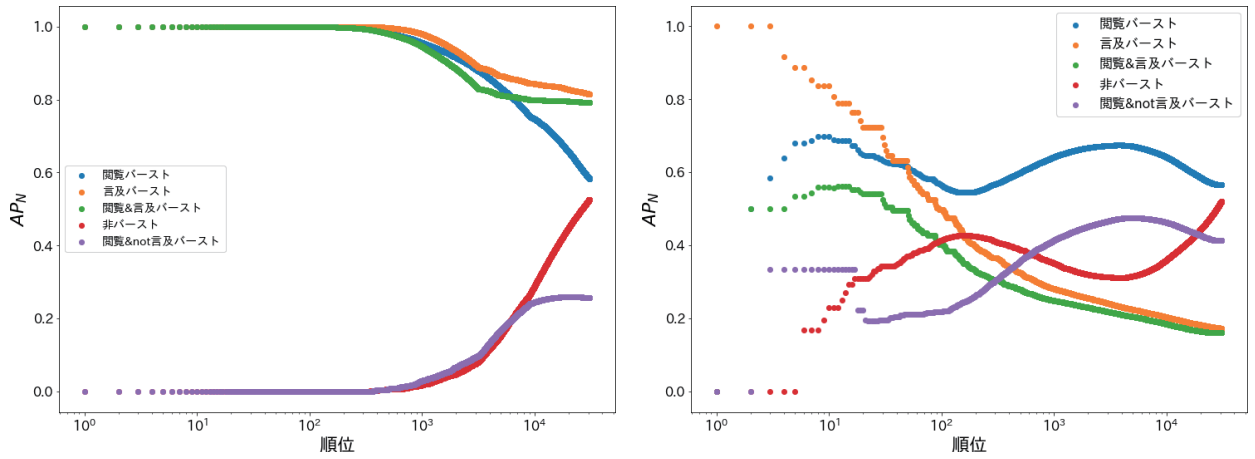
次に，閲覧・言及人数で評価した時の順位とバーストの関係について，**平均適合率**（Average Precision）を用いて分析する．平均適合率は，情報検索や情報推薦などの分野で複数の正解がある順位付きの結果を評価するための指標であり，例えば N 個の順位付きの結果に対する平均適合率 AP_N は，次の式で求められる．

$$AP_N = \sum_{k=1}^{N} \frac{P@k \cdot y_k}{\sum_{i=1}^{k} y_i} \tag{5}$$

ここで，$P@k$ は k 位までを対象にした時の適合率（Precision）であり，y_k は k 位が正解の場合は 1，そうでなければ 0 とする．平均適合率は，より多くの正解が，より上位にあるほど大きくなる．

言及・閲覧人数で順位付けした時のバースト論文の分布の違いを分析するために，横軸に順位を，縦軸にその順位（N）までの結果に対する AP_N をプロットした片対数グラフを図 1 に示す．

言及人数で順位付けした場合は，図 1(a) から，言及バースト論文のほとんどが上位に集中しており，言及バーストは言及人数の増加に大きく寄与することがわかった．また，言及と閲覧の両方でバーストした論文は，平均適合率の低下が早く言及でバースト論文の中で高い順位を占めているが，閲覧のみバーストした論文は下位になって平均適合率が増加しており，中程度の順位に分布していることがわかる．



(a) 言及人数　　(b) 閲覧人数

図 1 順位に伴うバースト論文の AP の関係

閲覧人数で順位付けした場合は，図 1(b) から，閲覧バースト論文の平均適合率には上位と下位で 2 つの山があり，前者は言及&閲覧バーストと連動し，後者は閲覧&not 言及バーストと連動していることがわかる．つまり，閲覧バーストには，ソーシャルメディアの言及由来のバーストと，別の要因のバーストの 2 種類が存在し，それぞれ閲覧人数の順位の異なる範囲に分布していると考えられる．

5.3 バースト論文の特性比較分析

そこで，閲覧バースト論文を，言及バーストの有無で 2 種類に分けて，論文題名や検索語を用いて比較する．閲覧人数上位の論文の詳細の閲覧・言及人数の順位，題名，検索語（バースト検索語は下線），バースト検索語が論文の著者名かどうか（1 または 0）を，表 2 に示す．

言及バーストした場合は，表 2(a) を見ると，2 位や 5 位の「下着」,「コーヒーカップ」のように一般的な興味を引くようなトピックや，7 位や 8 位の「伊野尾慧」,「小保方」などの有名人や実世界の事件に関する論文が出現し，論文の研究的価値というより，論文のトピックや著者への興味で閲覧されていると考えられる．

言及バーストしなかった場合は，表 2(b) を見ると，3 位の看護研究の文献検索や 21 位の深層学習に関する解説論文が出現し，これらは研究や教育目的で閲覧されていると考えられる．

5.4 閲覧バーストの期間と強さの特性分析

閲覧バースト論文を，言及バーストの有無で 2 種類に分けて，論文のバーストの期間と回数の関係に違いがあるか分析する．図 2 に横軸に平均バースト期間，縦軸にバースト回数をプロットしたグラフを示す．言及バーストした場合は，図 2(a) を見るとバースト回数が多く，言及バーストしなかった場合は，図 2(b) を見ると平均バースト期間が長いことがわかった．

長期的に閲覧バーストする論文の例として，図 3 に本研究で用いた論文特性分析システムでの出力を示す．調査の結果，この論文は，出版日（2014-07-25）とほぼ同時に，研究会発表（2014-08-02）で受賞したために，学術的に注目されて閲覧バーストしていることがわかった．これらの結果から，バーストの日時や期間，回数に着目することで，ネット上の言及や学術的なイベント由来でバーストする論文を識別できると考えられる．

6 おわりに

本稿では，オルトメトリクスで用いられるソーシャルメディア上の論文言及数と論文検索サービス上の論文閲覧数などの論文評価指標の特性を，その時系列表現のバーストに着目して分析した．その結果，バースト論文は各指標が高くなりやすい傾向や，言及バーストは閲覧バーストも引き起こしやすいが，言及由来ではないバーストも存在し，それぞれ論文を閲覧した目的や，バーストの期間や数が異なることを示した．

今後は，本研究をさらに進展させて，例えば研究者や一般人などのユーザの種別に応じた論文検索・推薦を実現するための技術について検討する予定である．

謝辞

本研究は JSPS 科研費 19H04421 の助成を受けた．

参考文献

[1] 林和弘. 科学技術動向研究 研究論文の影響度を測定する新しい動き: 論文単位で即時かつ多面的な測定を可能とする Altmetrics. 科学技術動向, pp. 20–29, 2013.

[2] Jason Priem, Heather A. Piwowar, and Bradley M. Hemminger. Altmetrics in the wild: Using social media to explore scholarly impact. arXiv:1203.4745 [cs.DL], 2012.

[3] 佐藤翔, 吉田光男. 日本の学協会誌掲載論文のオルトメトリクス付与状況. 情報知識学会誌, Vol. 27, No. 1, pp.

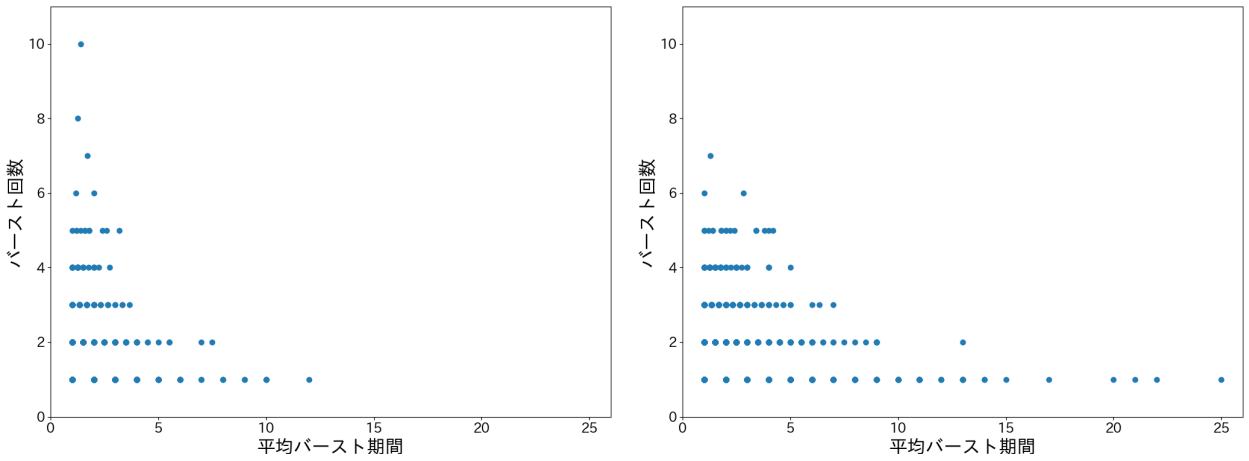
表 2 閲覧人数上位の閲覧バースト論文の詳細

(a) 言及バーストした場合

閲覧人数順位	言及人数順位	題名	検索語	バースト著者
2	3	ブラジャー着用時と非着用時の運動中の乳房振動特性	ブラジャー,乳房, 振動, 特性,ブラ	0
4	40	ポスドクからポストポスドクへ	ポスドク,円城塔, 円城, ポストポスドク, 問題	0
5	1	コーヒーカップとスプーンの接触音の音程変化	コーヒーカップ,スプーン, インスタントコーヒー,コーヒー, 塚本浩司	0
7	4	7336 津波避難に係る学校施設の整備のあり方：津波被害のあ...	伊野尾慧,伊野尾,大学院,明治大学, 論文	1
8	2288	海外情報 Harvard Medical School での再生医療教育	小保方晴子,小保方,晴子,harvard,medical	1

(b) 言及バーストしていない場合

閲覧人数順位	言及人数順位	題名	検索語	バースト著者
3	9148	看護のための文献検索のポイント：医中誌 Web を使って	検索, 文献,看護研究, 看護,無料	0
18	2288	高機能自閉症児は健常児と異なる「心の理論」をもつのか：「誤っ...	自閉症,別府哲,高機能自閉症,誤信,発達心理学	0
21	4885	画像認識のための深層学習 (< 連載解説 >Deep Learning...	深層,deep learning, 学習,連載,deep	0
22	9148	スパースコーディングの基礎理論と画像処理への応用	スパースコーディング, スパース・コーディング,スパース, スパースモデル,画像処理	0
26	2288	ポリアクリル酸ナトリウムの毒性試験の概要	ポリアクリル酸ナトリウム, ポリアクリル,na,安全性,アクリル酸	0



(a) 言及バーストした場合　(b) 言及バーストしていない場合

図 2 平均バースト期間とバースト回数の関係

図 3 長期的に閲覧バーストする論文例

23–42, 2017.

[4] 佐藤翔, 石橋柚香, 南谷涼香, 奥田麻友, 保志育世, 吉田光男. Twitter からの言及数が多い論文は言及されたことのない論文と比べてタイトルが「面白い」. 情報知識学会誌, Vol. 29, No. 3, pp. 268–283, 2019.

[5] Ehsan Mohammadi, Mike Thelwall, Mary Kwasny, and Kristi L. Holmes. Academic information on Twitter: A user survey. *PLOS ONE*, Vol. 13, No. 5, p. e0197265, May 2018.

[6] Jon Kleinberg. Bursty and hierarchical structure in streams. *In Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 91–101, 2002.

[7] 佐藤由将, 大竹恒平, 生田目崇. Twitter 上のバースト現象とコンテンツとの関係分析ーテレビドラマを例としてー. 日本ソーシャルデータサイエンス学会論文誌, Vol. 2, pp. 32–38, 2018.

[8] 佐々木謙太朗, 田村一樹, 吉川大弘, 古橋武. Twitter における話題語の抽出と周期に基づく分類. 言語処理学会第 19 回年次大会, pp. 806–809, 2013.

[9] 吉田光男. 計量書誌学の新たな挑戦：国産オルトメトリクス計測サービスの開発 (<特集>計量書誌学を超えて). 情報の科学と技術, Vol. 64, No. 12, pp. 501–507, 2014.

[10] 佐藤翔. コンテンツ入手元として機関リポジトリが果たしている役割. PhD thesis, 筑波大学, 3 2013.

[11] Liu Bing, editor. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, chapter Web Usage Mining. Springer-Verlag, 2009.

ARG WI2 No.16, 2020
WI2-2020-19

Twitter 上の arXiv プレプリントに関する学術情報流通のキーパーソンの特性分析

嶋田 恭助[†, a]　風間 一洋[†, b]　吉田 光男[††, c]　大向 一輝[†††, d]
佐藤 翔[††††, e]　桂井 麻里衣[†††††, f]

† 和歌山大学大学院システム工学研究科　†† 豊橋技術科学大学情報・知能工学系
††† 東京大学大学院人文社会系研究科　†††† 同志社大学 免許資格課程センター　††††† 同志社大学理工学部

a) *s216327@wakayama-u.ac.jp*　b) *kazama@wakayama-u.ac.jp*　c) *yoshida@cs.tut.ac.jp*　d) *i2k@l.u-tokyo.ac.jp*
e) *min2fly@slis.doshisha.ac.jp*　f) *katsurai@mm.doshisha.ac.jp*

概要　本稿では，プレプリントサービスが世の中に与えている影響を分析するために，arXiv 論文の情報拡散者のツイートと情報収集者のリツイートやお気に入りを 2 部グラフでモデル化し，学術情報流通におけるキーパーソンの関係と役割を分析する．そのために，情報収集者集合の類似性から作成した情報拡散者の関連ネットワークからコミュニティ抽出し，Kleinberg の HITS アルゴリズムのオーソリティ度を情報拡散者の重要度とする．まず，論文の分野から，大部分が人工知能関連であり，特に最大コミュニティは深層学習のキーパーソンを多数含むことを示す．さらに，Twitter のプロフィール言語とコミュニケーション言語から最大コミュニティは英語でも，関連コミュニティはコミュニケーション言語が異なり，プロフィール言語との不一致も多いことから，国際コミュニティと地域コミュニティを橋渡しするキーパーソンの存在を示す．

キーワード　Twitter，arXiv，学術情報流通，HITS アルゴリズム，2 部グラフ

1　はじめに

近年のオープンサイエンスへの移行により，学術論文が広く一般公開される機会が増加しているが，いまだにレベルの高い論文は少数の商業出版社による寡占状態にあり [1]，必ずしも論文の全文を自由に閲覧できない．そこで，査読前の論文や最新の研究成果の公開を目的としたプレプリントサーバが研究情報の交換や学会の運営に積極的に活用されている．特に COVID-19 の世界的な流行に対処するために，医療分野を中心にプレプリントサーバは重要な役目を果たしていると言える [2][1]．

研究者の情報交換も，学会開催時などの限られた機会だけでなく，Twitter などのソーシャルメディアで随時議論できるようになり，学術情報もこれらのサービス上を流通するようになった [3]．つまり，重要なプレプリントは該当分野の専門的なユーザにより発見・告知され，それを読んだユーザも重要だと判断した場合にいいねやリツイートを繰り返し，広く拡散すると考えられる．

ただし，Twitter 上の学術情報流通の解析には，いくつかの困難な課題が存在する．まず，公式リツイートでは正確な情報流通経路はわからない．次に，ユーザが持つ情報拡散者と情報収集者の二面性を考慮する必要がある．さらに，arXiv は新しいプレプリントを告知する公式 bot を提供しており，フォロワーも多いことから，従来のように単純に bot を除去せずに，人間と bot を包括して分析する必要がある．

本稿では，プレプリントサービスが世の中に与えている影響を分析するために，ソーシャルメディア上の学術情報流通の発信源として重要な役目を果たしているキーパーソンの関係と役割を分析する．まず，arXiv プレプリントに関する Twitter 上のツイート，リツイート，いいねの 3 種類の行動から，ユーザをノードとするグラフ構造を作成する．次に，ユーザには情報の拡散と収集の 2 種類の役割があるとみなして，その相補的な関係を HITS アルゴリズムで分析して，高いオーソリティ度を持つユーザを学術情報流通のキーパーソンとみなす．さらに，正のオーソリティ度を持つ情報拡散者の関係ネットワークから Louvain 法でコミュニティを抽出し，各コミュニティにおいてキーパーソンがどのような役割を果たしているかを，媒介中心性とユーザのプロフィールと情報交換に使う言語情報，ネットワーク可視化を用いて分析して明らかにする．

2　関連研究

arXiv プレプリントやソーシャルメディアにおける情報共有に関する研究が存在する．三根は，1991 年から 2007 年までの arXiv プレプリントを分析した結果，47.1%は学術雑誌に掲載され，特に高エネルギー物理学分野は arXiv 登録時点で既に掲載済みであり，arXiv がプレプリントサーバとしての役割の他に，研究内容の一般公開を目的とした学術情報メディアの役目を持つことを示した [4]．ただし，データの期間から宇宙物理学，物性物

Copyright is held by the author(s).
The article has been published without reviewing.
[1] https://rcos.nii.ac.jp/miho/2020/09/20200905/

理学，数学が中心で，本稿で扱っている 2013 年以降に大幅に急増した人工知能分野は分析されていない．

吉田らは，Twitter 上の arXiv プレプリントに対するツイート，いいね，リツイート，メンションの 4 種類の行動タイプの基礎調査を行った結果，ある行動をするユーザは他の行動をあまり行わない傾向があり，複数の行動タイプを同時に持つことは少なかったとしている [5]．すなわち，本稿で対象とする情報収集者と情報拡散者は，自然発生的に役割分担していると仮定できる．

3 学術情報流通の分析手法

3.1 ユーザの重要度の抽出

Twitter 上の学術情報流通という観点から，ユーザの情報拡散者と情報収集者という異なる重要度を抽出する．

ユーザ $u_i (i = 1, \ldots, N)$ は，情報拡散者 u_i^s と情報収集者 u_i^c としての側面を持つとして，情報収集者 u_i^c が情報拡散者 u_j^s のツイートをリツイートやいいねした場合に，u_j^s から u_i^c に情報が伝播したと考え，その伝播の有無 $D_{i,j}$ を要素とする N 行 N 列の隣接行列 $\boldsymbol{D}$ で表す．

Twitter 上では，有益と考えられる情報拡散者は多くの信頼できる情報収集者によってリツイートやいいねされ，その逆も成り立っている情報拡散者と情報収集者の 2 部グラフ構造が構築されていると考えられる．そこで，Kleinberg の **HITS**(Hyperlink-Induced Topic Search) アルゴリズム [6] を用いて，ユーザ u_i の情報拡散者及び情報収集者としての重要度 (a_i, h_i) を次の手順で求める．

1. 各ユーザの情報拡散者としての重要度を表すオーソリティベクトル $\boldsymbol{a} = (a_1, \ldots, a_N)^\top$，情報収集者としての重要度を表すハブベクトル $\boldsymbol{h} = (h_1, \ldots, h_N)^\top$ は，隣接行列 $\boldsymbol{D}$ を用いて式 (1) と式 (2) を適用後に正規化することを値が収束するまで繰り返す．

$$\boldsymbol{a} = \boldsymbol{D}^\top \boldsymbol{h} \quad (1)$$

$$\boldsymbol{h} = \boldsymbol{D}\boldsymbol{a} \quad (2)$$

   なお，$\boldsymbol{a}$ と $\boldsymbol{h}$ の各要素の初期値は 1 とする．
2. この結果から，ユーザ u_i の**オーソリティ度**と**ハブ度**の組 (a_i, h_i) を作成する．

3.2 情報拡散者のコミュニティの抽出

情報拡散者のコミュニティを，情報拡散者の関係ネットワークから次の手順で抽出する．

1. 各情報拡散者 u_i^s にリツイート・いいねした情報収集者の集合 $U_i^c = \{u_j^c | d_{j,i} = 1\}$ を作成する．
2. 情報拡散者 u_i^s と u_j^s の Simpson 係数 $simpson(U_i^c, U_j^c)$ が閾値 T 以上の場合にエッジ $e_{i,j}$ を張り，情報拡散者の関係ネットワーク G^s を作成する．
3. G^s を Louvain 法 [7] でコミュニティ $C_1^s, \ldots, C_K^s$ に分割する．

3.3 ユーザの言語特性の抽出

Twitter 上では，通常のツイートやリプライなどのコミュニケーションには，自分あるいは所属する母語を用いる．これを**コミュニケーション言語**と呼ぶ．これに対して，例えば，国際的な学術論文の発表には，英語などの母語とは異なる言語を用いることがある．このような仕事や学業の影響は自分を紹介するプロフィールに現れると考えて，プロフィールの記述言語も抽出する．これを**プロフィール言語**と呼ぶ．ユーザの Twitter 上の言語特性は，この 2 種類の言語情報の組として扱う．

コミュニケーション言語は，Twitter API を用いて得られる lang フィールドを用いる．プロフィール言語は，プロフィールのテキストから URL 除去後に，Python の langdetect[2] ライブラリで判定する．言語は，ISO 639-1 の 2 文字のアルファベットで表す．

4 分析

4.1 データセット

2007 年 3 月 21 日から 2020 年 1 月 18 日までの間に arXiv プレプリントについて呟いたツイートを Twiter API で収集し，ツイート中の圧縮 URL を展開してから，arXiv プレプリントの**言及データセット**として使用した．ただし，Twitter API にはツイートに関するリツイートといいねを最大 100 件までしか取得できない制限があるが，今回のデータセットでは，該当した事例はいいねは 5,600 件，リツイートは 1,449 件しかなかったので，大きな問題にはならない．

さらに，2020 年 2 月 26 日に arXiv から OAI-PMH [3] を用いて 1,645,129 本の論文の書誌情報を収集して，**arXiv メタデータセット**として使用した．各論文の書誌情報は，論文 ID，著者名（複数可），投稿日，更新日，タイトル，カテゴリ（主・副あり，複数可），抄録，コメント，DOI，Journal reference，Report number，ACM class，MSC class を含む．カテゴリは，計算機科学（cs）などの 8 種類の収録アーカイブ名を大分類とし，さらに詳細な分類の研究分野をピリオドで結合した文字列（例，人工知能は cs.AI）である．

4.2 言及ツイートの分析

まず，言及データセットの詳細を，表 1 に示す．言及ツイートに対して，いいねは約 26%，リツイートは約 14%であることから，いいねは気軽におこなっても，リツイートでは内容を評価・選別している可能性が高い．

[2] https://pypi.org/project/langdetect/
[3] http://www.openarchives.org/OAI/openarchivesprotocol.html

表 1: 言及データセットの詳細

項目	値
ツイート総数	3,088,669
言及された論文数	981,865
ユーザ総数	586,999
言及したユーザ数	118,743
いいねされたツイート数	797,294
リツイートされたツイート数	446,652

表 2: 重要度の詳細

a オーソリティ度・ハブ度

項目	値
オーソリティ度 >0	64,490
ハブ度 >0	566,367
オーソリティ度，ハブ度 >0	43,858
オーソリティ度 >0 & ハブ度 = 0	20,632
ハブ度 >0 & オーソリティ度 = 0	522,509
オーソリティ度の最大値	0.007356
ハブ度の最大値	0.0004568839

b 入次数・出次数

項目	値
入次数 >0	64,490
出次数 >0	566,367
入次数 , 出次数 >0	43,858
入次数 >0 & 出次数 = 0	20,632
出次数 >0 & 入次数 = 0	522,509
入次数の最大値	14,134
出次数の最大値	8,519

次に，arXiv プレプリントの言及特性を分析するために，論文の言及人数とユーザの言及論文数の分布を図 1 に示す．横軸に順位を，縦軸に言及人数または言及論文数を取り，両対数プロットした．言及人数に関しては，突出した 1 位の論文を除けば，図 1a に示すようにほぼ直線状で，べき分布であると考えられる．言及論文数に関しては，図 1b に示すように，約 10^2 から約 10^4 の急激な増加があり，その両端はほぼ直線状であった．高順位のユーザの言及論文数は約 $10^4 \sim 10^5$ とかなり多く，bot である可能性が高いと考えられる．

4.3 ユーザの重要度の特徴分析

ユーザの情報拡散者・情報収集者としての HITS のオーソリティ度・ハブ度の特性と，入次数・出次数との違いを分析する．オーソリティ度・ハブ度と入次数・出次数の詳細を表 2 に示す．表 2a を見ると，オーソリティ度が正のユーザは 11.0%しか存在せず，学術情報流通に関わるユーザの中でも情報拡散者はごく一部に限られ，ほとんどが情報収集しかしていないことがわかる．表 2b を見ると，入次数・出次数も値は異なるが同じ傾向を示す．

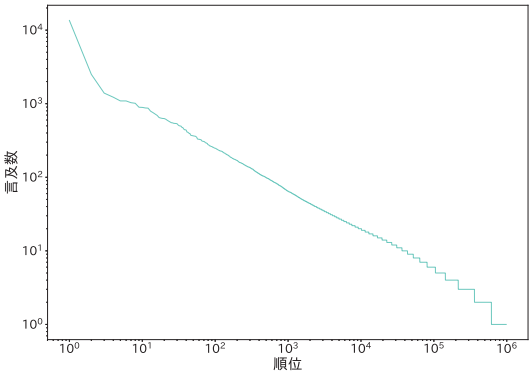
次に，オーソリティ度・ハブ度，入次数・出次数の上位ユーザの違いについて分析する．括弧内の数字は，組となる別の指標の順位である．なお，学術情報に一度も言及していない場合は，プロフィールを取得しないために，ユーザ名が空白となる．表 3a と表 3c に示すように，オーソリティ度と入次数の上位には，公式 bot や著名研究者，著名企業が含まれることがわかる．これに対して，表 3b と表 3d に示すように，ハブ度と出次数の上位には，そもそも自分から情報拡散していないユーザが多いことがわかる．また，対応する指標であるオーソリティ度と入次数，ハブ度と出次数では，上位のランキングはかなり異なることがわかる．

スピアマンの順位相関係数 ρ を計算すると，オーソリティと入次数は 0.999 と高い相関が，ハブと出次数は 0.541 と中程度の相関があることがわかった．

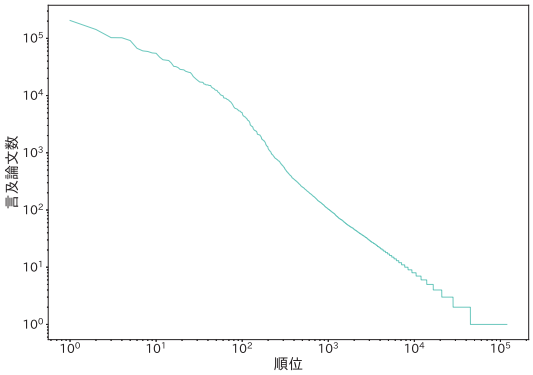
4.4 情報拡散者のコミュニティの分析

次に情報拡散者だけに注目し，そのコミュニティ構造を分析する．Simpson 係数の閾値は 0.5 として抽出した結果，ノード数 5,034，エッジ数 20,119 の情報拡散者ネットワークとなり，その最大連結成分から 25 個のコミュニティが得られた．図 2 に，最大連結成分の情報拡散者ネットワークのコミュニティ構造を示す．Gephi の ForceAtlas2 でレイアウトした後に，ノードサイズはオーソリティの大きさ，エッジの太さは Simpson 係数に応じて変化させ，コミュニティ別に着色した．図 2 左上にある巨大なノード集合（コミュニティ7,8,4,13）は機械学習系のコミュニティであり，右下のノード集合（コミュニティ0,2,6）は物理系のコミュニティである．また，コミュニティ7 は日本の機械学習系コミュニティでコミュニティ1 は日本の物理コミュニティとなっている．さらに，ユーザ数の上位 10 コミュニティの arXiv カテゴリを表 5 に示す．情報拡散カテゴリは言及した arXiv 論文の，情報収集カテゴリはいいね・リツイートした論文のカテゴリである．これから，特に物理系と機械学習系の情報拡散者が多いことがわかる．

次に情報拡散者ネットワーク中でキーパーソンがどこに位置するかと，評価指標としてオーソリティ度と入次数がどのように違うかを分析する．図 2 と同じレイアウト結果に対して，ノードだけを描画すると同時に，オーソリティ度・入次数の小さいノードほど透明度が高くなるように描画し，エッジは無視して可視化した結果を，図 3 に示す．この結果から，どちらもキーパーソンと思われるノードが各クラスタに分布しているが，図 3b の入次数を用いた場合には近接するノードでも色の濃さに大きな差があるのに対して，図 3a に示すオーソリティ度を用いた場合にはノードが密集している部分は全体的に色の濃さが近くなり，逆にノードが楚な部分は色が薄くなっていることがわかる．これは，HITS アルゴリズムは基本的に次数ベースの手法なので得られる結果の類似性は高くても，特に多くのオーソリティとハブによる 2 部グラフ的な構造が存在する場合にオーソリティ度・ハブ度共に値が大きくなる特性を持つからであり，本稿



a 言及人数



b 言及論文数

図 1: 言及人数と言及論文数の分布

表 3: 上位 10 人のユーザの詳細

a オーソリティ度

	ユーザ名	オーソリティ度	ハブ度
1	hardmaru	0.007356	0.000295 (25)
2	Miles Brundage	0.006711	0.000219 (79)
3	arxiv	0.005886	0.000000 (574486)
4	DeepMind	0.005608	0.000004 (50110)
5	Stat.ML Papers	0.004696	0.000000 (575126)
6	Ian Goodfellow	0.003937	0.000095 (995)
7	Alex J. Champandard	0.003858	0.000094 (1013)
8	Tomasz Malisiewicz	0.003768	0.000068 (1860)
9	Reza Zadeh	0.003619	0.000018 (10989)
10	Thomas	0.003618	0.000193 (132)

b ハブ度

	ユーザ名	オーソリティ度	ハブ度
1		0.000000(181109)	0.000457
2		0.000000(584335)	0.000424
3	priya joseph	0.000000(103529)	0.000412
4	fly51fly	0.000033(3944)	0.000402
5		0.000000(503464)	0.000385
6	Vincent Boucher ?	0.000069(2285)	0.000384
7	AssistedEvolution	0.000019(5816)	0.000381
8	Hamid EBZD	0.000253(727)	0.000376
9	PerthMLGroup	0.000008(11480)	0.000370
10	MONTREAL.AI	0.000581(259)	0.000358

c 入次数

	ユーザ名	入次数	出次数
1	hardmaru	14134	611 (24)
2	arxiv	13645	0 (582413)
3	DeepMind	12475	7 (44513)
4	Miles Brundage	11569	425 (63)
5	Xavi Bros	10515	28 (10008)
6	Daisuke Okanohara	9293	7 (45820)
7	Stat.ML Papers	7259	0 (568010)
8	Alex J. Champandard	7176	122 (1024)
9	Reza Zadeh	6908	28 (9910)
10	Alessandro Vespignani	6672	57 (3715)

d 出次数

	ユーザ名	入次数	出次数
1		0(181109)	8519
2		0(503464)	6285
3	Xinder App	22(11066)	1594
4		0(584335)	1384
5	priya joseph	0(96930)	1167
6	Hamid EBZD	173(2014)	1036
7		0(562674)	1001
8	Igor Carron	59(5590)	983
9	Vincent Boucher ?	53(6029)	922
10	fly51fly	27(9718)	879

のようにキーパーソン同士が協調して情報流通に貢献しているような構造を分析したい場合には，基本的に個々のノードに対する指標であるとともに，大きなコミュニティほど大きくなる特性を持つ入次数よりも，オーソリティ度の方が適していると考えられる．

5 情報拡散者のコミュニティの言語特性分析

情報拡散者のコミュニティの言語特性について分析する．表5に，各コミュニティのコミュニケーション言語とプロフィール言語のユーザ数上位 3 件を降順で示す．表中にある und は不明言語を意味し，言語特定ができなかったものである．

ユーザ数が 1 番多い言語に着目すると，一番巨大なコミュニティ8 や 4 はコミュニケーション言語，プロフィール言語共に英語（en）であるのに対し，コミュニティ1 や 7 のように共に日本語（ja）である小さなコミュニティが存在することがわかる．ただし，そのような場合でも 2 番目は共に英語であり，その割合も少なくはない．また，この表には示していないユーザ数が 6 のコミュニティ105 は，一番多いコミュニケーション言語は韓国語でも，プロフィール言語は英語の方が多いという逆転現象が生じている．つまり，例えば同じ機械学習分野でも，複数のコミュニティに分割される理由は，言語または地域的な違いが原因であると考えられる．また，以上のことから，コミュニティ8 や 4 は国際的な活動をしているコミュニティで，コミュニティ1 や 7 は地域的な活動をしているコミュニティでないかと推測できる．実際に，表 3a で示したオーソリティ度の上位の著名研究者や著名企業などのキーパーソンはコミュニティ8 や 4 に属していた．

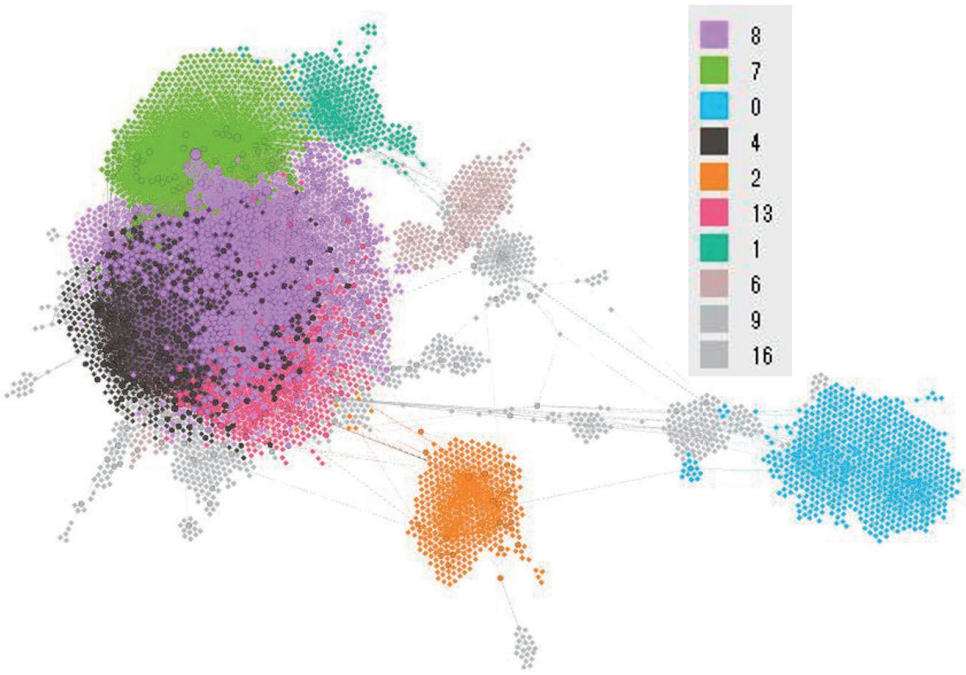


図 2: 情報拡散者ネットワークの可視化結果

表 4: コミュニティごとの情報拡散・収集カテゴリ

コミュニティ	ユーザ数	情報拡散カテゴリ	情報収集カテゴリ
0	628	astro-ph.EP：196，astro-ph.GA：165	astro-ph.EP：211，astro-ph.GA：198
1	256	hep-th：43，quant-ph：29	hep-th：46，quant-ph：33
2	377	physics.soc-ph：177，cs.SI：47	physics.soc-ph：226，cs.SI：38
4	606	cs.LG：311，cs.CV：99	cs.LG：392，cs.CV：77
6	245	quant-ph：115，physics.chem-ph：30	quant-ph：141，physics.chem-ph：35
7	685	cs.LG：206，cs.CV：174	cs.LG：310，cs.CV：160
8	1414	cs.LG：651，cs.CV：249	cs.LG：836，cs.CV：175
9	184	hep-ph：52，hep-ex：46	hep-ex：55，hep-ph：52
13	314	cs.CL：255，cs.LG：32	cs.CL：260，cs.LG：40
16	119	math.CT：23，cs.LO：20	math.CT：35，cs.PL：20

さらにコミュニケーション言語とプロフィール言語に注目すると，コミュニティ1や7であってもプロフィール言語の方が英語の情報拡散者数が多く，日本語は少なくなっている．これは，国際的な活動をしているためにプロフィールは英語で書くが，地域的なコミュニケーションや貢献のために母語でコミュニケーションする機会も多いからではないかと考えられる．つまり，地域的なコミュニティには，国際的なコミュニティとの橋渡しをしている別の種類のキーパーソンが存在する可能性がある．

そこで，どのようなユーザがコミュニティ間の橋渡しをしているかを調べるために，媒介中心性の上位10ユーザのコミュニケーション言語（clang）とプロフィール言語（plang），コミュニティ（com），媒介中心性（bc）の値を表6に示す．媒介中心性ではオーソリティ度が高いbotやコミュニティ4に所属するIan Goodfellow（オーソリティ度6位），Alex J. Champandard（オーソリティ度7位）のような著名なキーパーソンは存在しないものの，やはり国際的なコミュニティ8，4に所属するキーパーソンが多かった．さらに，地域的なコミュニティ7に属するDaisuke Okanohara（オーソリティ度17位）はコミュニケーション言語とプロフィール言語が異なっており，さらに20位は同じコミュニティで同じの言語特性のYuta Kashino（オーソリティ度68位）であった．地域的なコミュニティにはオーソリティ度も高く，媒介中心性も高いユーザが確認された．

6 おわりに

本稿では，arXiv論文の情報拡散者のツイートと情報収集者のリツイートやお気に入りを2部グラフでモデル化し，学術情報流通の情報拡散におけるキーパーソンの関係や役割を，重要度，コミュニティ，言語特性の観点から分析した．

arXivカテゴリから，大部分のコミュニティが物理学か

a オーソリティ度　　b 入次数

図 3: キーパーソンの可視化結果

表 5: コミュニティのコミュニケーション言語・プロフィール言語

コミュニティ	ユーザ数	コミュニケーション言語	プロフィール言語
0	628	en : 583, und : 5, es : 4	en : 546, und : 32, it : 8
1	256	ja : 182, en : 42, und : 10	ja : 154, en : 62, ko : 11
2	377	en : 357, tl : 2, ca : 1	en : 326, und : 19, es : 4
4	606	en : 568, und : 8, pt : 4	en : 528, und : 33, de : 12
6	245	en : 233, und : 2, ja : 2	en : 213, und : 17, fr : 3
7	685	ja : 506, en : 121, und : 22	ja : 346, en : 218, und : 26
8	1414	en : 1341, und : 19, fr : 2	en : 1244, und : 63, de : 21
9	184	en : 165, es : 9, und : 3	en : 154, es : 11, und : 10
13	314	en : 302, und : 6, ja : 1	en : 276, und : 18, de : 10
16	119	en : 111, ro : 2, es : 2	en : 91, und : 14, de : 2

表 6: 媒介中心性上位 10 ユーザ

ユーザ名	clang	plang	com	bc
hardmaru	en	en	8	0.249
Miles Brundage	en	en	8	0.226
Kyle Cranmer	en	en	9	0.171
Daisuke Okanohara	ja	en	7	0.142
Gianfranco Bertone	en	en	9	0.128
? FDN Intelligence	en	en	4	0.111
Alessandro Vespignani	en	en	2	0.091
Benjamin Weiner	en	en	0	0.088
Pietro Vischia	en	en	9	0.048
Olof Nebrin	en	en	0	0.045

人工知能関連であり，特に最大の英語でコミュケーションしている国際的なコミュニティには，オーソリティ度が高い深層学習のキーパーソンを多数含むことを示した．これに対して，主に母語でコミュニケーションすることが多い地域的なコミュニティには，プロフィール言語が英語でも，コミュニケーション言語が母語の媒介中心性が高い，別の種類のキーパーソンが存在することを示した．

今後はさらに別の特徴も加えて分析すると共に，本研究をさらに発展させて，学術情報流通分析に基づいた新しい論文評価指標を検討する予定である．

謝辞

本研究は JSPS 科研費 19H04421 の助成を受けた．

参考文献

[1] Vincent Larivière, Stefanie Haustein, and Philippe Mongeon. The Oligopoly of Academic Publishers in the Digital Era. *PLOS ONE*, Vol. 10, No. 6, pp. 1–15, June 2015.

[2] 小柴等, 林和弘, 伊藤裕子. COVID-19 / SARS-CoV-2 関連のプレプリントを用いた研究動向の試行的分析. Technical report, 科学技術・学術政策研究所, 2020.

[3] Andrea Chiarelli, Rob Johnson, Stephen Pinfield, and Emma Richens. Accelerating scholarly communication: The transformative role of preprints. *Zenodo*, September 2019.

[4] 三根慎二. 学術情報メディアとしての arXiv の位置づけ. *Library and information science*, No. 61, pp. 25–58, 2009.

[5] 吉田光男, 嶋田恭助, 風間一洋, 佐藤翔. arXiv 論文に対する Twitter での言及行動タイプに関する予備調査. 人工知能学会第 34 回全国大会, 2020.

[6] Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *J. ACM*, Vol. 46, No. 5, pp. 604–632, September 1999.

[7] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2008, No. 10, p. P10008, October 2008.

ARG WI2 No.16, 2020
WI2-2020-20

Twitterでの学術情報流通におけるネットワーク作成法とユーザ重要度の関係

豊島 秀典　　吉田 光男　　梅村 恭司
豊橋技術科学大学
toyoshima.hidenori.xi@tut.jp yoshida@cs.tut.ac.jp umemura@tut.jp

概要　学術情報流通の分析では，ソーシャルネットワーク上のユーザに対して重要度を示す指標が用いられる．一方で，学術情報流通において，ネットワーク作成方法の違いによる，指標の変化やユーザアクティビティとの関係は知られていない．本研究では，論文言及への反応を用いたネットワークとフォロワを用いたネットワークによる，HITS スコアの変化を調べた．さらに，それぞれの HITS スコアと学術情報に関するユーザアクティビティの関係について調べた．それらの結果，高い HITS オーソリティ度をもつユーザの HITS スコアはネットワークの違いにより，大きく変化することが分かった．また，両ネットワークとユーザアクティビティとの関係がある程度一致していることが分かった．これらの結果は，フォロワを用いたネットワークと言及への反応を用いたネットワークのいずれにおいても，HITS スコアがある意味での重要度を表していることを示している．

キーワード　Twitter，arXiv，学術情報流通，HITS アルゴリズム

1　はじめに

学術情報流通の分析では，ソーシャルネットワーク上のユーザに対して重要度を示す PageRank [1] やベクトル中心性 [2], HITS [3] などの指標が用いられる [4, 5, 6]．一般に Twitter を分析する際，フォロワを用いたネットワークが用いられる [7]．一方で，フォロワを用いたネットワークと比べ，リツイートなど反応を用いたネットワークがよりユーザ間の実世界での関係を捉えているという報告もある [8]．

ソーシャルネットワーク上でのネットワークの違いが分析される一方で，学術情報流通においては，ネットワーク作成方法の違いによる，ユーザ重要度指標の変化や，学術論文に関するユーザアクティビティとの関係は知られていない．ネットワーク作成方法は処理する情報の選別を行うものであるため，その検討は分析に重要と考えられる．また，処理する方法が異なれば結果が異なることは容易に想像できるものの，どのように異なるかを想像するのは難しい．本研究では，学術情報流通における，ネットワーク作成方法の違いが与える，ユーザ重要度指標の変化と，ユーザアクティビティとの関係を調べる．

2　分析方法

学術情報流通を分析するに当たって，本研究では，学術情報への言及の多いソーシャルネットワークである Twitter[1] と著名なプレプリントサーバ arXiv[2] を対象とする．本稿では，反応を用いたネットワークと，フォロワを用いたネットワークの 2 種類の学術情報流通ネットワークを比較する．これらのネットワークに対してユーザの重要度をそれぞれ計算し，ユーザ重要度の上位ユーザを比較する．また，ユーザアクティビティと重要度の関係を調べる．ユーザアクティビティには言及論文数と，arXiv への投稿日からツイート日までの日数平均を用いる．

2.1　学術情報流通ネットワーク

arXiv 論文へのリンクを含むツイート（言及ツイート）を投稿したユーザの集合を情報拡散者集合 U^t とする．情報拡散者集合に対して，情報を収集したユーザ集合を情報収集者集合 U^c とする．学術情報流通ネットワークは，情報拡散者と情報収集者との関係を表すグラフである．

ユーザ $u_i \in U^t$ から情報を得たユーザ集合を，ユーザ u_i の情報収集者集合 $U_i^c \subseteq U^c$ とする．$N = |U^t \cup U^c|$ に対して，情報流通のネットワークを N 行 N 列の以下のような行列 D で表す．

$$D = \begin{bmatrix} d_{1,1} & d_{1,2} & \dots & d_{1,N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N,1} & d_{N,2} & \dots & d_{N,N} \end{bmatrix} \quad (1)$$

$u_i, u_j \in U^t \cup U^c (1 \leq i, j \leq N)$ に対して $u_i \in U^t$ かつ $u_j \in U_i^c$ であった時，u_j が u_i の情報を得たとし，$d_{i,j} = 1$ とする．

2.2　反応を用いたネットワーク

反応を用いたネットワークとして，嶋田らの提案した学術情報流通の 3 層モデルを準用する．詳細については，文献 [6] も参考にされたい．

反応を用いたネットワークでは，言及ツイートに対し

Copyright is held by the author(s).
The article has been published without reviewing.
[1] https://twitter.com
[2] https://arxiv.org

て，いいね・リツイート・返信を行ったユーザ集合 U^r を情報収集者集合 U^c とする．また，$u_i \in U^t$ のツイートに対して，反応をしたユーザ集合を U_i^r とし，ユーザ u_i の情報収集者集合 U_i^c とする．小節 2.1 の手順に従い，U^t と U^r から作成した行列 D を，反応を用いたネットワーク D^r とする．学術情報に対する言及行動では，ユーザによって取り得る行動の種類が異なる [9] ことから，いいね・リツイート・返信のすべてを反応として扱うことで，多くのユーザの行動を反映したグラフを作成できる．

2.3 フォロワを用いたネットワーク

フォロワを用いたネットワークでは，言及ツイートを行ったユーザのフォロワ集合 U^f を情報収集者集合 U^c とする．また，$u_i \in U^t$ のフォロワ集合を U_i^f とし，ユーザ u_i の情報収集者集合 U_i^c とする．小節 2.1 の手順に従い，U^t と U^f から作成した行列 D を，フォロワを用いたネットワーク D^f とする．

2.4 ユーザ重要度の計算

ユーザの重要度として HITS [3] のオーソリティ度を用いる．小節 2.2，2.3 で作成した，反応を用いたネットワーク D^r とフォロワを用いたネットワーク D^f を用いて，HITS のオーソリティ度をそれぞれ計算する．

3 データ

分析には，学術情報流通のツイートデータと，学術情報データを用いる．データの作成方法は文献 [6] と同様である．

3.1 学術情報流通のツイートデータ

Twitter Search API を用いて arXiv 論文へのリンクを含むツイートを収集し，2007 年 3 月 21 日から 2020 年 1 月 18 日までのツイート 3,077,409 件を取得した．ここでは，文字列 “arxiv.org” と arXiv ID を含むツイートを言及ツイートとした．また，言及ツイートに対して反応（いいね・リツイート・返信）したユーザと，言及ツイートを行ったユーザのフォロワをそれぞれ取得した．ただし API の仕様により，いいね・リツイート・返信に関しては，ツイートに対してそれぞれ 100 ユーザ程度までしか取得できていない場合がある．

言及を行ったユーザ数 $|U^t|$ は 118,743 であり，言及へ反応したユーザを含めたユーザ数 $|U^t \cup U^r|$ は 609,121，フォロワを含めたユーザ数 $|U^t \cup U^f|$ は 8,339,328 である．

小節 2.2 にしたがって作成したネットワークは，ノード数 609,121，エッジ数 2,032,829 である．また，出次数 0 のユーザ数は 23,144 であり，入次数 0 のユーザは 541,215 である．小節 2.3 にしたがって作成したネットワークは，ノード数 8,339,328，エッジ数 7,683,307 である．また，出次数 0 のユーザ数は 656,021 であり，入次数 0 のユーザは 8,257,895 である．2 つのネットワークで，入次数がともに 1 以上のユーザ数は 50,873 である．

3.2 学術情報データ

arXiv プレプリントサーバから OAI-PMH を用いて学術情報を取得した．本研究では，論文が投稿された日付を用いた．3.1 小節で取得した言及ツイートで言及された論文数は 981,865 である．

4 結果と考察

本節では，2 種類のネットワークを用いた HITS オーソリティ度の分析結果と考察を示す．

4.1 HITS のオーソリティ度上位ユーザ

オーソリティ度上位ユーザを比較することで，ネットワークの違いにより，出現するユーザが異なるかどうかを調べる．また，ユーザのアカウントを確認することで，Bot などの特徴的なユーザが上位に集中していないかどうかも確かめる．

反応を用いたネットワークとフォロワを用いたネットワークについて，オーソリティ度上位 20 ユーザを表 1 にまとめた．@arxiv_org と @StatMLPapers は Bot とみられる．@hardmaru，@DeepMind，@goodfellow_ian ら，7 アカウント（表中下線）が双方の上位 20 ユーザにも出現し，26 アカウントが一方のみに出現した．また，反応を用いたネットワークとフォロワを用いたネットワークについて，オーソリティ度が 0 を超えるユーザを取り出し，スピアマンの順位相関係数を求めた．50,873 ユーザに対する順位相関係数は 0.09 であり，ほとんど相関はみられなかった．

反応を用いたネットワークとフォロワを用いたネットワークで，求まるユーザの重要度は異なる傾向を持つことが分かる．どちらのネットワークを用いても，重要なユーザと考えられるものが上位に現れており，これをさらに分析することで重要という観点を細分化できる可能性を示唆している．

4.2 HITS のオーソリティ度ヒストグラム

オーソリティ度のユーザ分布を確認し，ユーザの分布に偏りが存在するかどうかを確かめる．

図 1a に反応（リツイート・いいね・返信）を用いた HITS のオーソリティ度のヒストグラム，図 1b にフォロワを用いた HITS のオーソリティ度のヒストグラムを示す．縦軸はユーザ数，横軸はオーソリティ度を示し，右図と左図は同じデータをビンの幅を変えて出力している．

図 1a のオーソリティ度 10^{-5} 付近の狭い範囲にユーザが集中している．このことから，反応（リツイート・いいね・返信）を用いた HITS のオーソリティ度の分布はフォローを用いた HITS のオーソリティ度の分布に比べて偏りがあることが分かる．

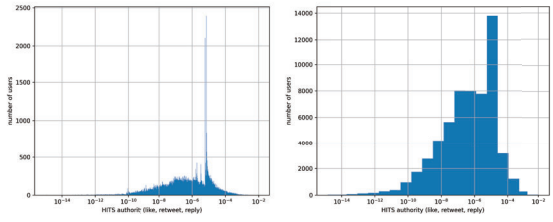
表 1 HITS オーソリティ度上位ユーザ

a. 反応を用いたネットワーク

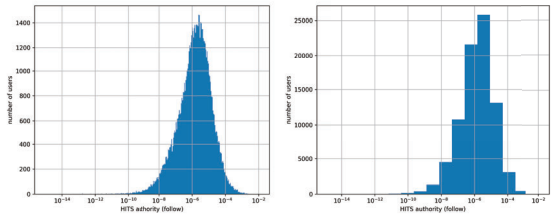
	表示名	ユーザ名
1	hardmaru	@hardmaru
2	Miles Brundage	@Miles_Brundage
3	arxiv	@arxiv_org
4	DeepMind	@DeepMind
5	Stat.ML Papers	@StatMLPapers
6	Ian Goodfellow	@goodfellow_ian
7	Alex J. Champandard	@alexjc
8	Tomasz Malisiewicz	@quantombone
9	Reza Zadeh	@Reza_Zadeh
10	Thomas Lahore	@evolvingstuff
11	ML Review	@ml_review
12	Andrej Karpathy	@karpathy
13	François Chollet	@fchollet
14	samim	@samim
15	Russ Salakhutdinov	@rsalakhu
16	Nenad Tomasev	@weballergy
17	Daisuke Okanohara	@hillbig
18	Oriol Vinyals	@OriolVinyalsML
19	Denny Britz	@dennybritz
20	Nando de Freitas	@NandoDF

b. フォロワを用いたネットワーク

	表示名	ユーザ名
1	Andrew Ng	@AndrewYNg
2	Ian Goodfellow	@goodfellow_ian
3	DeepMind	@DeepMind
4	François Chollet	@fchollet
5	Fei-Fei Li	@drfeifei
6	OpenAI	@OpenAI
7	Demis Hassabis	@demishassabis
8	Jeff Dean	@JeffDean
9	Hugo Larochelle	@hugo_larochelle
10	Oriol Vinyals	@OriolVinyalsML
11	Russ Salakhutdinov	@rsalakhu
12	Nando de Freitas	@NandoDF
13	hardmaru	@hardmaru
14	Soumith Chintala	@soumithchintala
15	Jeremy Howard	@jeremyphoward
16	Stanford NLP Group	@stanfordnlp
17	TensorFlow	@TensorFlow
18	Christopher Manning	@chrmanning
19	Chris Olah	@ch402
20	Richard Socher	@RichardSocher



(a) 反応（リツイート・いいね・返信）



(b) フォロワ

図 1 HITS のオーソリティ度のヒストグラム

ユーザの反応は，API の仕様により 1 ツイートに対して 100 ユーザ程度までしか取得できない制限がある．一方，フォロワのネットワークにこのような制限は存在しない．そのためこの制限が，反応を用いた HITS のオーソリティ度の分布に偏りが生じる一因であると考える．

4.3 投稿からツイートまでの平均日数

投稿からツイートまでの平均日数を比較し，情報を早く発信することと，オーソリティ度に関係があるか調べる．

図 2a に反応（リツイート・いいね・返信）を用いたグラフ，図 2b にフォロワを用いたグラフを示す．縦軸は投稿からツイートまでの平均日数を，横軸はオーソリティ度を示す．左図は実数グラフ，右図は両対数グラフである．

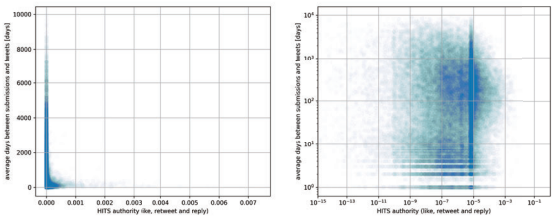
図 2a のオーソリティ度 10^{-5} 付近に縦の筋が存在する．図 1a から，図 2a のオーソリティ度 10^{-5} 付近に存在する縦の筋はユーザ数がオーソリティ度 10^{-5} 付近に偏って存在しているためであると分かる．また，両対数グラフにおいて，図 2a，2b のグラフの概形が一致していることが分かる．さらに，両対数グラフにおいて，右上の角が切れていることから，重要度が高いユーザに反応の遅いユーザが存在しないことが分かる．

0 を超えたオーソリティ度と，投稿からツイートまでの平均日数に対して，スピアマンの順位相関を求めた．反応のネットワークでは 67,906 ユーザで，相関係数は −0.38，フォロワのネットワークでは 81,433 ユーザで −0.23 であり，共に負の相関が見られた．順位相関から，いずれのネットワークでも，arXiv 投稿に早く反応・言及したユーザがより高い重要度を持つ傾向にあることが分かる．

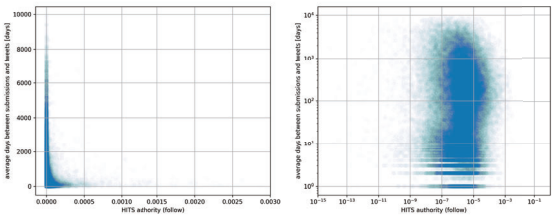
4.4 arXiv 論文への言及量

arXiv 論文への言及ツイートでの言及論文数の総和について比較し，arXiv 論文への言及量，つまり情報の発信量とオーソリティ度に関係があるかどうかを調べる．

図 3a に反応（リツイート・いいね・返信）を用いたグラフを，図 3b にフォロワを用いたグラフを示す．縦

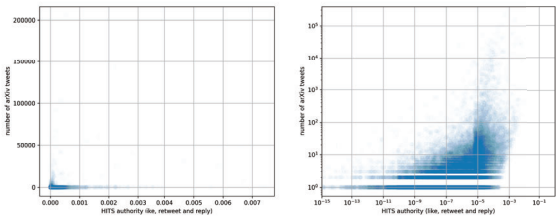


(a) 反応（リツイート・いいね・返信）

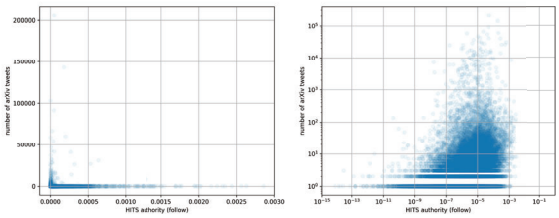


(b) フォロワ

図 2 HITS のオーソリティ度と反応日数



(a) 反応（リツイート・いいね・返信）



(b) フォロワ

図 3 HITS のオーソリティ度と arXiv 言及ツイート数

軸は言及論文数，横軸はオーソリティ度を示す．左図は実数グラフ，右図は両対数グラフである．

図 3a，3b において，両対数グラフにおいて左上の角が切れていることから，言及量が少ないユーザに重要度が高いユーザはいないことが分かる．また，両対数グラフにおいて，図 3a，3b のグラフの概形が一致していることが分かる．さらに，反応のネットワークでは，フォロワのネットワークに比べ，言及数の多いユーザの集団がオーソリティ度が高い方向へ移動していることから，反応のネットワークの方がフォロワのネットワークに比べ，言及量が重要度に影響を与えることが分かる．

0 を超えたオーソリティ度と，arXiv 論文への言及量に対して，スピアマンの順位相関を求めた．反応のネットワークでは 67,906 ユーザで，相関係数は 0.50，フォロワのネットワークでは 81,433 ユーザで 0.33 であり，共に正の相関が見られた．順位相関から，いずれのネットワークでも，言及量の多いユーザがより高い重要度を持つ傾向にあることが分かる．

5　おわりに

本論文では，論文言及への反応を用いたネットワークとフォロワを用いたネットワークによる，HITS スコアの変化を調べた．また，それぞれの HITS スコアと，言及量などの学術情報に関するユーザアクティビティの関係について調べた．その結果，高い HITS オーソリティ度をもつユーザの HITS スコアはネットワークの違いにより，大きく変化することが分かった．また，両ネットワークとユーザアクティビティとの関係がある程度一致していることが分かった．これらの結果は，フォロワを用いたネットワークと言及への反応を用いたネットワークのいずれにおいても，HITS スコアがある意味での重要度を表していることが分かった．

参考文献

[1] Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank citation ranking: Bringing order to the web., Technical report, Stanford InfoLab (1999).

[2] Bonacich, P.: Power and centrality: A family of measures, *American journal of sociology*, Vol. 92, No. 5, pp. 1170–1182 (1987).

[3] Kleinberg, J. M.: Authoritative sources in a hyperlinked environment, *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632 (1999).

[4] Ke, Q., Ahn, Y.-Y. and Sugimoto, C. R.: A systematic identification and analysis of scientists on Twitter, *PloS one*, Vol. 12, No. 4, e0175368 (2017).

[5] Hoffmann, C. P., Lutz, C. and Meckel, M.: A relational altmetric? Network centrality on R esearch G ate as an indicator of scientific impact, *Journal of the Association for Information Science and Technology*, Vol. 67, No. 4, pp. 765–775 (2016).

[6] 嶋田恭助，風間一洋，吉田光男，佐藤翔：Twitter 上の arXiv からの学術情報流通に関する分析，第 12 回データ工学と情報マネジメントに関するフォーラム，No. A7-5 (2020).

[7] Myers, S. A., Sharma, A., Gupta, P. and Lin, J.: Information network or social network? The structure of the Twitter follow graph, *Proceedings of the 23rd International Conference on World Wide Web*, pp. 493–498 (2014).

[8] Bild, D. R., Liu, Y., Dick, R. P., Mao, Z. M. and Wallach, D. S.: Aggregate characterization of user behavior in Twitter and analysis of the retweet graph, *ACM Transactions on Internet Technology*, Vol. 15, No. 1, Article 4 (2015).

[9] 吉田光男，嶋田恭助，風間一洋，佐藤翔：arXiv 論文に対する Twitter での言及行動タイプに関する予備調査，2020 年度人工知能学会全国大会，No. 1L4-GS-5-02 (2020).

ARG WI2 No.16, 2020
WI2-2020-21

Random Walk with Restart とコサイン類似度に基づく研究者推薦モデル

中村 幹太[a]　　岡本 一志[b]

電気通信大学 大学院情報理工学研究科 情報学専攻

a) *k.nakamura@uec.ac.jp*　b) *kazushi@uec.ac.jp*

概要　本研究では，Random Walk with Restart と研究内容のコサイン類似度に基づいた，有向グラフによる研究者推薦モデルを提案する．有向グラフは科学研究費助成事業データベースから構築し，過去の共同研究関係，研究内容のコサイン類似度，およびそれらを組み合わせた 3 種類のエッジを定義し，それぞれを history モデル，similarity モデル，combination モデルとする．3 つのモデルごとに推薦精度として nDCG@k を比較する．実験結果として，過去の共同研究関係とコサイン類似度を組み合わせた combination モデルの nDCG@k が最も高くなることを確認している．また，ネットワーク構造の比較により，エッジ数の増加だけでなく次数の分布も推薦精度に影響することを確認している．

キーワード　共同研究者推薦，有向グラフ，学術データベース，ランダムウォーク

1　はじめに

適切な共同研究者の発見は共同研究の成功に寄与するだけでなく，研究者同士でアイデアや知識を交換することで新たな知見を得ることができる．そのため共同研究者推薦が試みられており，大きく 2 つのモデルに分類できる．ひとつは過去の共同研究関係による研究者ネットワークの構造情報を使用する構造情報型モデルで，もうひとつは公開された論文や特許から取得される研究内容のテキストデータを使用する内容型モデルである．構造情報型モデルは同分野の研究者推薦に有効であり，内容型モデルは異分野の研究者推薦に有効とされている [1]．構造情報型の代表的な推薦モデルとして，Random Walk with Restart（RWR）がある．これはネットワークの構造情報から，ノードの類似度を算出し推薦するモデルである．研究者ネットワークは一般的なソーシャルネットワークよりも密度が低いと考えられ，共同研究関係以外のエッジを定義することで推薦精度を向上できる可能性がある．また，異分野などでエッジが繋がっていない研究者は推薦できない．

本研究では，RWR にコサイン類似度に基づくエッジ導入した研究者推薦モデルを提案する．このようにすることでネットワークの密度が向上し，ネットワーク上の距離が遠い研究者も推薦できるようになると考える．共同研究数をエッジとした history モデルをベースラインとして，研究内容の類似度をエッジとする similality モデル，それらを組み合わせた combination モデルを提案する．実験では構築した推薦モデルによって推薦リストを出力し，推薦精度として normalized Discounted Cumulative Gain(nDCG)@k を比較する．

Copyright is held by the author(s).
The article has been published without reviewing.

2　関連研究

構造情報型の研究者推薦に用いられる RWR は，ノード u からのランダムウォークによる各ノードへの遷移確率を定量化するモデルで，ノード同士の類似度とみなされる [2]．i 回ランダムウォークしたときのノード u から各ノードへの遷移確率 $\mathbf{r}_i$ は

$$\mathbf{r}_i = cA^{\mathrm{T}}\mathbf{r}_{i-1} + (1-c)\mathbf{q}_u, \tag{1}$$

のように表される．ここで $\mathbf{q}_u$ および $\mathbf{r}_0$ は u に対応する要素を 1，他の要素を 0 とした one-hot 列ベクトルである．隣接行列 A は各行ベクトルの総和が 1 になるように正規化されており，$A_{u,v}$ はノード u から v への遷移確率を表す．c は隣接ノードへの遷移確率を表し，$1-c$ は開始ノード u に戻る確率である．i 回のランダムウォーク後，$\mathbf{r}_i$ はノード u と他のノード間の類似度とみなされ，研究者推薦に利用されている [3, 4, 5]．

Li らは RWR を応用した研究者推薦モデルを提案している [4]．このモデルでは共同研究数と共同研究が開始されてからの期間によって隣接行列が重み付けされている．推薦精度は DBLP（digital bibliography&library project）データセットで評価され，RWR モデルよりも推薦精度が高いことを示している．Kong らは RWR によって計算された類似度を特徴量とした，教師あり学習による研究者推薦モデルを提案している [5]．DBLP データセットで推薦精度を評価し，RWR モデルおよび Li らの推薦モデル [5] の precision よりも高くなることを示している．

Araki らは研究内容のコサイン類似度に基づいた研究者推薦モデルを提案している [1]．推薦精度は KAKEN データセットで評価され，異分野に対する推薦精度が Kong らのモデルよりも高いことを示している．また，

異分野に対する研究者推薦では構造情報型モデルよりも内容型モデルが有効であることを示している．

Nakamura らは有向研究者ネットワークを使用した教師あり学習型の研究者推薦モデルを提案している [6]．有効研究者ネットワークによって研究代表者と分担者の関係を表現し，研究の役割を考慮した研究者推薦を行なっている．提案モデルでは研究者ベクトルの外積･結合･重み付き和をエッジベクトルとして，ロジスティック回帰によってリンク成立確率を予測し，推薦している．推薦精度を KAKEN データセットで評価し，研究者ベクトルの外積の推薦精度が最も高いことを示している．Araki らおよび Nakamura らの結果より，研究者ベクトルのコサイン類似度や外積は研究者推薦にとって重要な指標であると考える．

3 RWR とコサイン類似度基づく研究者推薦

RWR における異分野間の研究者推薦では内容型情報を用いる研究者推薦よりも推薦精度が低い [1]．これは研究者ネットワークにおける異分野間のエッジが少ないことに起因していると考えられる．既存の RWR モデル [4, 5] は共同研究関係の重み付け方法を提案しているが，研究者ネットワークのエッジを追加する研究はなく，推薦精度が向上する可能性がある．そこで本研究では，RWR とコサイン類似度に基づくエッジを使用した研究者モデルを提案する．

3.1 コサイン類似度に基づくエッジの定義

提案モデルでは研究者 u に推薦スコアに基づいた上位 k 人の研究者を推薦する．本研究ではノード $u, v \in V$，エッジ $y_{u,v} = \{0, 1\} \in Y$ として研究者ネットワークを有向グラフ $G = \{V, Y\}$ として定義する．ここで $y_{u,v}$ は研究者 u, v 間にエッジが存在する場合 1，存在しない場合 0 となる．また，それぞれの研究課題には 1 人の研究代表者と 1 人以上の研究分担者が存在すると想定している．ここでベースラインモデルとなる history エッジ，提案モデルとなる similarity エッジを示す．

1. history エッジ
   研究分担者 u と研究代表者 v の過去の共同研究の集合 $W_{u,v}$ を用いて，history エッジ $e_h(u, v)$ を

$$e_h(u, v) = |W_{u,v}| \tag{2}$$

   と定義する．

2. similarity エッジ
   研究者 u の研究内容の特徴ベクトルを $\mathbf{x}_u$ として，研究分担者 u と研究代表者 v 間の similarity エッジ $e_s(u, v)$ を

$$e_s(u, v) = \begin{cases} \frac{\mathbf{x}_u^{\mathrm{T}} \cdot \mathbf{x}_v}{||\mathbf{x}_u|| \cdot ||\mathbf{x}_v||} & u \neq v \\ 0 & u = v \end{cases} \tag{3}$$

   と定義する．

3.2 隣接行列の構築

エッジ関数 e_h, e_s に対して隣接行列をそれぞれ

$$A_h = \begin{pmatrix} \frac{e_h(1,1)}{\sum_{i=1}^{|V|} e_h(1,i)} & \cdots & \frac{e_h(1,|V|)}{\sum_{i=1}^{|V|} e_h(1,i)} \\ \vdots & \ddots & \vdots \\ \frac{e_h(|V|,1)}{\sum_{i=1}^{|V|} e_h(|V|,i)} & \cdots & \frac{e_h(|V|,|V|)}{\sum_{i=1}^{|V|} e_h(|V|,i)} \end{pmatrix} \tag{4}$$

$$A_s = \begin{pmatrix} \frac{e_s(1,1)}{\sum_{i=1}^{|V|} e_s(1,i)} & \cdots & \frac{e_s(1,|V|)}{\sum_{i=1}^{|V|} e_s(1,i)} \\ \vdots & \ddots & \vdots \\ \frac{e_s(|V|,1)}{\sum_{i=1}^{|V|} e_s(|V|,i)} & \cdots & \frac{e_s(|V|,|V|)}{\sum_{i=1}^{|V|} e_s(|V|,i)} \end{pmatrix} \tag{5}$$

と定義する．隣接行列 A_h, A_s は各行ベクトルの総和が 1 になるよう正規化する．また，A_s は計算効率の観点から閾値を設定し，エッジを削除する．e_s が t 未満のエッジを削除する閾値 t と，各行の e_s 上位 n 件以外を削除する閾値 n を定義する．A_h と A_s を組み合わせた combination モデルの隣接行列 A_c を

$$A_c = \frac{1}{2}(A_h + A_s) \tag{6}$$

と定義する．

3.3 推薦リスト

推薦スコアは隣接行列 A を用いた式（1）を用いて計算する．得られた $\mathbf{r}_i$ を推薦スコアの降順でソートし，上位 k 人の研究者を推薦リストとして出力する．

なお，イテレーション i は無限大であることが望ましく，そのときの $\mathbf{r_i}$ は

$$\lim_{i \to \infty} \mathbf{r}_i = (1 - c)(I - cA^{\mathrm{T}})^{-1}\mathbf{q}_u \tag{7}$$

で表される [7]．その場合 A の逆行列 $(I - cA^{\mathrm{T}})^{-1}$ は疎行列ではなく計算コストが高くなるため，本研究では式（1）を使用する．

4 研究者推薦精度の評価実験

提案モデルとベースラインモデルの推薦精度を KAKEN データセットを用いて評価する．

4.1 KAKEN データセット

KAKEN データセットの研究課題から有向研究者ネットワークを構築する．このデータセットを使用する理由

表 1 実験で用いる研究者ネットワークの概要

	開始年度	研究課題数	ノード数	エッジ数
探索	2001 – 2014	140,063	145,045	275,751
テスト	2001 – 2015	150,222	150,534	290,460

として，すべての研究分野の研究課題が含まれている点や，研究者が研究者番号によって一意に識別できる点が挙げられる．研究課題は研究課題番号，研究期間，研究者とその役割，タイトル，キーワード，および要約で構成されている．これらのデータは API を用いて収集し，2001 年度から 2017 年度に開始した研究課題を使用する．過去の共同研究関係は直近の共同研究関係に比べ，現在の共同研究に影響を与えないことが知られている [8]．計算効率の観点から，本研究では 2000 年度以前に開始した研究課題を使用しない．エッジは研究分担者から研究代表者方向のエッジのみが存在すると仮定する．

4.2 データセット分割とハイパーパラメータ探索

ハイパーパラメータの探索と推薦精度の比較のため，2 種類の研究者ネットワークを構築する．2001 年度から 2014 年度開始の研究課題から構築したネットワークを $G_{\mathrm{L}} = \{V_{\mathrm{L}}, Y_{\mathrm{L}}\}$ として，ハイパーパラメータの探索に用いる．また，2001 年度から 2015 年度開始の研究課題から構築したネットワークを $G_{\mathrm{V}} = \{V_{\mathrm{V}}, Y_{\mathrm{V}}\}$ として，テストに用いる．同様に 2001 年度から 2017 年度開始の研究課題から構築したネットワークを $G_{\mathrm{T}} = \{V_{\mathrm{T}}, Y_{\mathrm{T}}\}$ とする．実験で RWR を適用する研究者ネットワーク G_{L} および G_{V} の概要を表 1 に示す．ハイパーパラメータとして similarity モデルと combination モデルのエッジ e_s の値の閾値 $t \in \{0.2, 0.4, 0.6, 0.8\}$ および順位の閾値 $n \in \{10, 20, 30, 40, 50\}$ を決定する．ハイパーパラメータ探索時のデータセット分割方法を以下に示す．

1. G_{L} から $X_{\mathrm{L}} = \{u, v | u, v \in V_{\mathrm{L}}, y_{u,v} = 1\}$ 取得
2. G_{V} から $X'_{\mathrm{L}} = \{u, v | u, v \in V_{\mathrm{V}}, u, v \notin X_{\mathrm{L}}, y_{u,v} = 1\}$ 取得
3. X'_{L} から 500 件ランダムサンプリング，X''_{L} 取得
4. RWR を X_{L} に適用，X''_{L} で推薦精度算出

テスト時のデータセット分割方法を以下に示す．

1. G_{V} から $X_{\mathrm{V}} = \{u, v | u, v \in V_{\mathrm{V}}, y_{u,v} = 1\}$ 取得
2. G_{T} から $X'_{\mathrm{V}} = \{u, v | u, v \in V_{\mathrm{V}}, u, v \notin X_{\mathrm{V}}, y_{u,v} = 1\}$ 取得
3. RWR を X_V に適用，X'_{V} で推薦精度算出

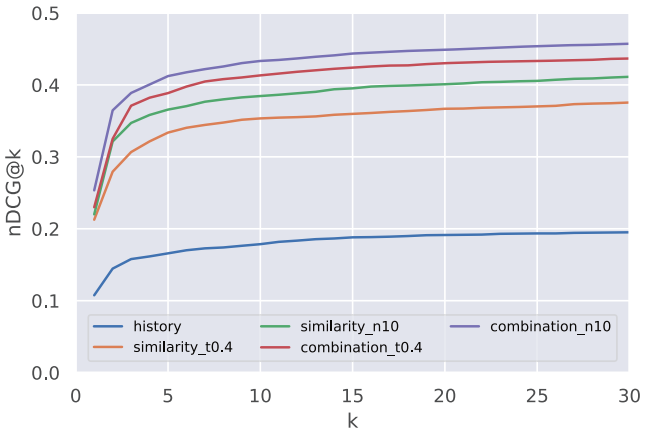


図 1 各モデルの平均 nDCG@k の比較

4.3 研究者ベクトルの構築と RWR

研究者 u の特徴ベクトル $\mathbf{x}_u$ の構築方法について説明する．形態素解析器である MeCab[9] を研究者 u が参画する研究課題のタイトル･キーワード･概要に適用し，普通名詞･固有名詞･サ行変格活用の名詞を抽出する．Term Frequency-Inverse Document Frequency（TF-IDF）を適用し，研究者ベクトル $\{\mathbf{x}_u | u \in V\}$ を取得する．

隣接行列 A_h, A_s, A_c を共同研究関係および研究者ベクトル $\mathbf{x}$ から構築する．研究者 u に対する RWR の計算コストは $O(|V|i)$ であり，研究者の数 $|V|$ とイテレーション i に依存する [7]．本実験ではイテレーション i を 100 に設定する．

4.4 推薦精度

研究分担者 u と実際に共同研究した研究代表者の集合を V_u^T とする．研究分担者 u に対する推薦リスト V_u^R の上位 j 番目の研究者を $v_j \in V_u^R$ として，DCG@k は

$$d = I(v_1 \in V_u^T) + \sum_{j=2}^{k} \frac{I(v_j \in V_u^T)}{\log_2 j} \tag{8}$$

のように表される．ここで，指示関数 $I(v_j \in V_u^T)$ は

$$I(v_j \in V_u^T) = \begin{cases} 1 & v_j \in V_u^T \\ 0 & v_j \notin V_u^T \end{cases} \tag{9}$$

を示す．nDCG@k は DCG@k を d の最大値 d_M で正規化した値であり，

$$d_n = \frac{d}{d_M} \tag{10}$$

のように表される．これらを研究者に関して平均したものを推薦精度として採用する．

5 研究者推薦精度の比較と考察

ハイパーパラメータ探索の結果，similarity モデルでは $t = 0.4$，$n = 10$ を選択している．また，combination モデルでも同様に $t = 0.4$，$n = 10$ を選択している．

表 2 history モデルと similarity モデルのネットワーク

モデル	ノード数	エッジ数	平均次数	次数の分散
history	1.51×10^5	2.90×10^5	1.93	3.83
similarity $t = 0.4$	1.51×10^5	3.68×10^6	24.4	4.59×10^3
similartiy $n = 10$	1.51×10^5	1.51×10^6	10	0

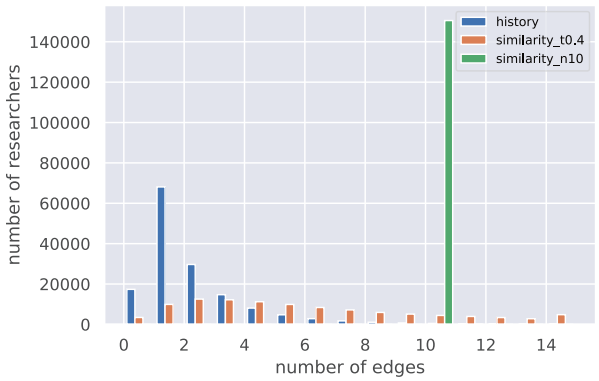


図 2 history モデルと similarity モデルの次数の分布

図 1 にテストデータにおける各モデルの平均 nDCG@k を示す．全ての提案モデルの nDCG@k が 0.1 以上 history モデルよりも高いことが確認できる．また，history モデルと similarity モデルを組み合わせた combination モデルの nDCG@k が最も高くなっていることが確認でき，異なる種類のエッジを組み合わせた場合も RWR が機能することを示している．また，similarity モデルおよび combination モデルいずれの場合も閾値 t よりも閾値 n の nDCG@k が高いことが確認できる．

表 2 に history モデルと similarity モデル $(t = 0.4)$ および $(n = 10)$ のネットワークの概要を示す．similarity モデル (n) は history モデルや similarity モデル (t) より nDCG@k は最も高いが，エッジ数は最も少ない．以上の点から，必ずしもエッジ数の増加が推薦精度の向上に寄与するとは限らない．次に history モデルと similarity モデル $(t = 0.4)$ および $(n = 10)$ の次数の分布を図 2 に示す．history モデルおよび similarity モデル (t) では次数が 0 のノードが存在することが確認できる．このような研究者には推薦できないため，similarity モデル (n) よりも nDCG@k が低いと考えられる．また，similarity モデル (n) の次数の分散が 0 である点も推薦精度に寄与している可能性がある．

6 おわりに

本研究では，有向研究者ネットワークのエッジの増加が RWR による研究者推薦精度を改善すると仮定し，推薦モデルにコサイン類似度に基づくエッジを導入している．計算コストの削減のため，コサイン類似度に対する閾値 t と次数に対する閾値 n を導入している．各モデルを比較した結果，history モデルと similarity モデルを組み合わせた combination モデルの nDCG@k が最も高いことを確認している．また，閾値 t モデルと n モデルでは n モデルの方が nDCG@k が高いという結果が得られており，エッジ数の増加だけでなく，その分散が小さいことが推薦精度の向上に寄与することを示唆している．

今後の研究として，新たなエッジを定義することや，combination モデルの各モデルを重み付けすることがある．また，推薦精度とネットワークの構造の関係性を明らかにすること，計算効率の改善を検討している．

謝辞

本研究は JSPS 科研費 JP18K18159 の助成を受けたものです．

参考文献

[1] M. Araki, M. Kstsurai, I. Ohmukai, H. Takeda, “Interdisciplinary Collaborator Recommendation Based on Research Content Similarity,” IEICE Transactions on Information and Systems, vol.E100.D, no.4, pp.785–792, 2017.

[2] J. Pan, H. Yang, C. Faloutsos, P. Duygulu, “Automatic Multimedia Cross-modal Correlation Discovery,” Proc. of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.653–658, 2004.

[3] H. Tong, C. Faloutsos, J. Pan, “Random Walk with Restart: Fast Solutions and Applications,” Knowledge and Information Systems, vol.14, pp.327–346, 2008.

[4] J. Li, F. Xia, W. Wang, Z. Chen, N. Asabere, H. Jiang, “ACRec: a Co-authorship Based Random Walk Model for Academic Collaboration Recommendation,” Proc. of 23rd International Conference on World Wide Web, pp.1209–1214, 2014.

[5] X. Kong, H. Jiang, Z. Yang, Z. Xu, F. Xia, A. Tolba, “Exploiting Publication Contents and Collaboration Networks for Collaborator Recommendation,” PLOS ONE, vol.11, no.2, pp.1–13, 2016.

[6] K. Nakamura, K. Okamoto, “Development of a Collaborator Recommender System Based on Directed Graph Model,” Proc. of 20th International Symposium on Advanced Intelligent Systems and International Conference on Biometrics and Kansei Engineering, pp.338–345, 2019.

[7] Y. Fujiwara, M. Nakatsuji, M. Onizuka, M. Kitsuregawa, “Fast and Exact Top-k Search for Random Walk with Restart,” Proc. of the VLDB Endowment, vol.5, no.5, pp.442–453, 2012.

[8] M. Hasan, V. Chaoji, S. Salem, M. Zaki: “Link prediction using supervised learning,” Proc of the SIAM Data Mining Workshop on Link Analysis Counterterrorism and Security, 2006.

[9] T. Kudo, K. Yamamoto, Y. Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis,” Proc. of the 2004 Conference on Empirical Methods in Natural Language Processing, pp.230–237, 2004.

ARG WI2 No.16, 2020
WI2-2020-22

性格要素と外見要素の加減算による類似キャラクタの検索

小林　達哉[a]　　松下　光範[b]
関西大学総合情報学部
a) *k252475@kansai-u.ac.jp*　b) *mat@res.kutc.kansai-u.ac.jp*

概要　本研究の目的は，ユーザの好みのキャラクタが登場するコミックを検索する技術の実現である．現在広く用いられているコミック検索では，作品名や著者名，ジャンル情報などコミックに付与された情報をクエリとしているが，コミックに登場するキャラクタの性格や外見といった情報は，コミックの選択に大きく影響するにも関わらず，それに基づく検索は十分に検討されていない．これを解決するための一助として，本稿ではキャラクタを表現する性格要素（e.g., 真面目，優しい）や外見要素（e.g., 金髪，長身）をクエリとして利用する検索手法を提案する．提案手法では，これらの要素を既知のキャラクタに加減算できるようにして好みのキャラクタを表現することで類似キャラクタの検索を可能にする．

キーワード　コミック，キャラクタ，検索，分散表現，IDF

1　はじめに

全国出版協会・出版科学研究所の調査 [5] によれば，コミックの新刊発行部数は毎年 1 万点以上にも及ぶ．日々増加を続けるこの膨大なコミックの中からユーザの興味や関心に合致した作品群に出会うのは困難であるため，そのような探索を支援するシステムの実現が求められている．

現状では，この膨大なコミックの中からユーザがコミックを検索しアクセスする手段として出版社のサイトや電子書籍販売サイト（e.g., コミックシーモア[1]，BookLive[2]）の検索機能が利用されることが多い．こうしたアクセスを行う際，ユーザはジャンル情報（e.g., 異世界，恋愛），作品名，著者名などをクエリとしてコミックを検索する．

しかし，これらのサービスではコミックに登場するキャラクタの性格や外見といった情報などは，コミックの選択に大きく影響するにも関わらず，キャタクタ情報に基づく検索は難しい．例えば，「熱血で優しいキャラクタが登場するコミック」や「天才で知的なキャラクタが登場するコミック」のような，内容に関わる情報をクエリとした検索は現状では行われていない．こうした問題点を解消するため，本研究では内容情報に基づくコミック検索の実現を試みる．

松井らはコミック検索支援システムの一つとして，コミックの登場キャラクタに着目した検索の手法を提案した [2]．この手法では，ユーザの既知のキャラクタを基に性格（e.g., 真面目，優しい）や外見（e.g., 金髪，短髪）要素を加減算することで「コナンに熱血を足したキャラクタ」のような検索を可能にしている．しかし，この研究ではキャラクタの要素を数値化する過程において，対象とするキャラクタがもつ性格や外見要素などを one-hot で表現しているためにそれらの間に重要度の差がなく，些細な特徴が影響して意図に沿った検索結果が必ずしも得られない．そこで，抽出された要素のうちどの要素がそのキャラクタを表現するうえで重要な要素かを反映可能にする必要がある．

こうした背景の下，本稿ではキャラクタを構成する要素として性格要素と外見要素に着目し，それぞれを分けて数値化することでこの問題の解決を図る．

Copyright is held by the author(s).
The article has been published without reviewing.

[1] https://www.cmoa.jp/
[2] https://booklive.jp/

2　先行研究

朴らはキャラクタ情報に基づくコミック検索を実現するために，Web 上から抽出したキャラクタの説明文を用いてエゴグラムによる性格分類を試みた [4]．この手法では東大式のエゴグラム [6] によりコミックのキャラクタの性格を五次元のベクトルで表現し，キャラクタの性格を推定した．被験者実験により，キャラクタの性格とユーザが認識するキャラクタの性格が一致するか検証したところ，55%の精度で推定できたことが報告されている．この結果は，キャラクタの性格に基づく類似度検索が可能であることを示しているが，現状では十分な精度とは言い難い．

松井らは性格情報だけでなくキャラクタの外見情報も必要であるとし，萌え要素の加減算に基づくキャラクタ検索手法を提案した [2]．この手法では，ユーザの既知のキャラクタを基に要素を追加・削除することで「コナンに熱血を足したキャラクタ」のような検索を可能にしている．しかし，この手法では性格や外見などの要素を区別せずに one-hot で数値化しているため，キャラクタを認識するうえで重要な特徴と些細な特徴が区別できず，検索者の意図に沿った検索結果が得られないことが

ある．

この問題を解決するため，本稿ではキャラクタの性格と外見を分けて数値化することで，抽出された要素のうちどの要素がキャラクタにとって重視されているかを表現可能にする．

3　デザイン指針

本稿では，既知のキャラクタを基に性格と外見の要素を加減算することによって，類似キャラクタの検索を可能にすることを目的とし，以下の 3 項目に取り組む．

（1）キャラクタの性格・外見に基づく数値化

（2）キャラクタの性格・外見の加減算に基づくキャラクタ検索

（3）キャラクタの演算を可能とするプロトタイプシステムの実装

(1）ではキャラクタを数値で表現するための性格・外見に基づく数値化の手法である．(2）では（1）の数値化手法に基づきキャラクタ要素の加減算を行う．これにより，例えば「キャラクタ A ＋キャラクタ B」の演算を行うことで，それぞれの特徴を持ち合わせたキャラクタを検索することが可能になる．また,「キャラクタ A ＋要素」を行うことにより,「コナンに熱血を足したキャラクタ」のように演算を行うことができる．(3）ではキャラクタ要素の加減算をするためのプロトタイプシステムを実装し，ユーザ実験を通じて提案手法の有効性や限界について検証する．

4　データセットの作成

データセットの作成は松井らの研究 [2] を参考に萌え要素に着目して行った．本稿では，萌え要素のうちキャラクタの構成要素である性格と外見要素を対象にした．処理手順は (1) 要素の収集, (2) 辞書の作成, (3) 要素の付与，の順に行った．各々の処理について以下に述べる．

4.1　要素の収集

萌え要素については，外見に基づく要素だけでなく性格の要素を含める [1]．これら萌え要素を表現する語は一個人の印象で確立されるものではなく，複数人の主観的な印象で構築される．そこで，要素のデータセットを構築するために複数のユーザが自由に書き込める自由参加型 Web 百科事典を対象とし要素を収集することにした．自由参加型 Web 百科事典はインターネットに接続することで誰もが記事を編集できる Web サイトである．本稿では Wikipedia[3], アニヲタ Wiki[4], ニコニコ大百科[5]，ピクシブ百科事典[6]のサイトを利用した．自由参加型 Web 百科事典では，萌え要素・萌え属性は同義語として扱われているため，各自由参加型 Web 百科事典で「萌え要素」または「萌え属性」として扱われている語を合計で 1403 語収集した．そこから重複する表現や公序良俗に反する表現を削除し，要素を表す 1172 語を対象とした．

ここで，萌え要素は性格と外見の要素が含まれているだけでなく，表現単体で特徴を表していないもの（e.g., うなじ，髪の色），キャラクタの性格と外見の特徴を表していない表現（e.g., 温泉回，世界観）がある．そのため，著者を含めて合計 7 人のアノテーション作業により，性格及び外見を端的に表現した単語（e.g., 明るい，冷静，金髪），短文（e.g., 世話好き，情が深い）のみを抽出した．抽出する条件として，4 人以上が性格または外見要素であると判断した場合その要素を抽出することとした．これにより 85 語の要素を削除し，1087 語をキャラクタの性格または外見の特徴を表している要素の表現として採用した．

4.2　辞書の作成

収集した要素を対象に，表 2 のように表記揺れの統一を行った．漢字・ひらがな・カタカナの表記揺れが存在する表現については，各表現ごとに weblio 辞書と自由参加型 Web 百科事典の記事を確認し，表記揺れとして記載されている表現を一つのカテゴリに統一した．例えば，統一前の「“タレ目，” “たれ目，” “垂目，” “垂れ目”」は意味的に同じであるが，漢字・ひらがな・カタカナの表記揺れにより複数の単語が存在することになるため，カタカナに統一した．このような統一処理を行うことで 1087 語の要素表現を 578 語（性格単語数：286 語，外見単語数：292 語）に絞り込んだ．

4.3　要素の付与

キャラクタに対して要素を付与するため，本稿では「漫画全巻ドットコム」[7]の歴代発行部数の中から 200 作品に登場する 1438 キャラクタを対象とし，自由参加型 Web 百科事典を用いてキャラクタの説明文が含まれた文章を収集した．取集した文章を形態素解析である MeCab[8]を用いることにより「名詞，形容詞」を抽出する．この時，固有名詞などを充実させた mecab-ipadic-neologd[9]の辞書を用いることによる表現の統一を行った．抽出された単語の中から本稿で作成した萌え要素の辞書に該当する単語をそのキャラクタに付与し，付与された一例を表 1

[3] https://ja.wikipedia.org/wiki/
[4] https://w.atwiki.jp/aniwotawiki/
[5] https://dic.nicovideo.jp/
[6] https://dic.pixiv.net/
[7] https://www.mangazenkan.com/ranking/books-circulation.html
[8] http://taku910.github.io/mecab/
[9] https://github.com/neologd/mecab-ipadic-neologd

表 1 キャラクタに要素を付与させた時の一例

キャラクタ名	うずまきナルト	孫悟空
性格	目立ちたがり，わがまま，冷静，明るい，負けず嫌い	明るい，穏やか，素直，優しい，純粋
外見	金髪，ゴーグル，ツンツン，碧眼，長身	筋肉質，ツンツン，糸目，小柄，尻尾

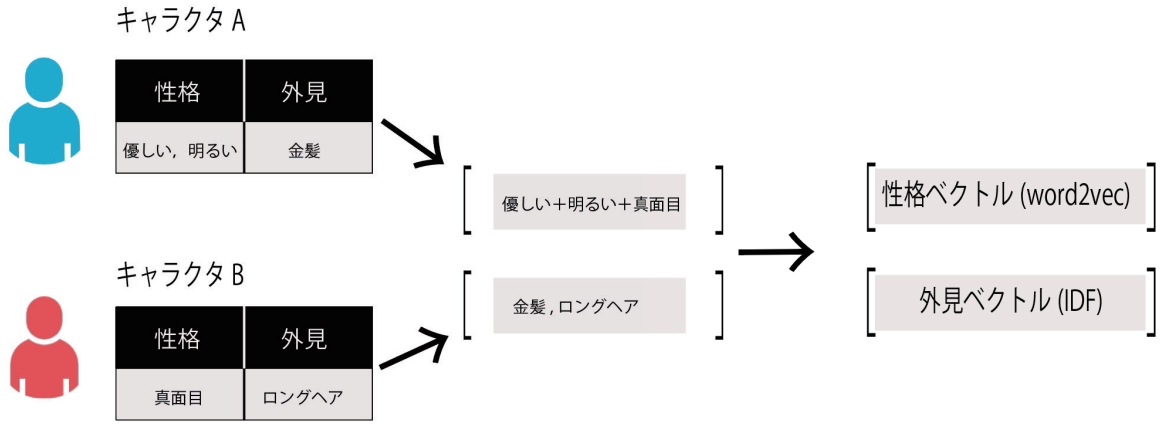


図 1 キャラクタ同士の演算によるベクトル作成

表 2 表記揺れ統一例（[2] より表引用）

統一前	統一後
タレ目，たれ目，垂目，垂れ目	タレ目
狐耳，きつね耳，キツネ耳	狐耳
ロングヘアー，ロングヘア，長髪	ロングヘア

に示す．

5 提案手法

本稿ではキャラクタを構成させる要素のうち性格と外見に着目して数値化を行い，要素を用いて加減算することで該当する別のキャラクタを検索する．加減算することで，「キャラクタ A に熱血を足したキャラクタ」や「キャラクタ A とキャラクタ B を足し合わせた」ような性格や外見に着目した類似キャラクタ検索を可能とする（図 1 参照）．

5.1 性格要素のベクトル化

キャラクタの性格は複数の側面を持ち合わせて構成されているため，加減算した単語の意味を複数考慮して性格を表現する必要がある．そこで，性格要素の意味を反映することができる単語の分散表現を用いて，キャラクタに単語ベクトルを加えることにより複数の性格の意味を反映させて性格を表現する．

以上の考えに基づき，キャラクタの性格要素は word2vec の手法により単語の分散表現でベクトルを獲得する．ここで，ベクトルの作成には Skip-gram[3] を用いた．これにより文章の語順に対応した学習が可能となり，周辺の単語分布が似ている単語ベクトルが似た値をとるようになる．ただし，キャラクタに付与した性格要素は辞書に該当する単語を抽出しているので語順の関係がないため，前後関係を考慮するような学習方法は不適切となる．そこで，語順学習をさせないため学習途中で性格要素の順序をランダムに変化させ語順の関係性を考慮させないよう学習した．word2vec の実装には python のライブラリである gensim（ver.3.8.3）[10]を用いた．パラメータは，次元数を 100 次元，ウィンドウサイズを 8 単語，学習 epoch 数を 100 に設定した．性格要素の語順をランダム化させる時は epoch 数を 5 毎にランダム化した．

cos 類似度に基づく学習単語の類似単語結果の一例を表 3 に示す．この結果を見ると，“明るい” は「“元気,” “天真爛漫,” “快活”」などの単語が高い類似度を示していることが確認できる．これらの語について weblio 辞書[11]で類似語検索を行い類語として分類されている語を確認したところ，入力単語である “明るい” という単語が類語として存在することが確認できた．

5.2 外見要素のベクトル化

キャラクタの外見は，場面や状況などのコミックの進行状況により衣装が変更されるように，外見要素の意味を考慮して構成されていない．そのため，要素の意味を反映して性格を表現する単語分散表現の手法は不適切である．そこで，キャラクタの外見を表現するために，対象とするキャラクタが持つ外見要素を数値化する．

[10]https://radimrehurek.com/gensim/
[11]https://thesaurus.weblio.jp/

表 3 単語の類似結果

単語	明るい		冷淡		真面目		ムードメーカー	
順位	類似単語	類似度	類似単語	類似度	類似単語	類似度	類似単語	類似度
1	元気	0.675	非情	0.754	堅物	0.706	人懐っこい	0.610
2	天真爛漫	0.672	無慈悲	0.647	頑固	0.566	明るい	0.542
3	快活	0.612	冷徹	0.606	優柔不断	0.507	快活	0.536
4	人懐っこい	0.604	冷静	0.552	律儀	0.481	明朗	0.537
5	活発	0.601	慎重	0.483	強情	0.464	男気	0.479

本稿では，キャラクタの外見要素に対して特徴語を考慮するために,「多数のキャラクタに登場する要素は,キャラクタの特徴語にはならない」という仮定に基づき,IDF 法を用いることとした．この手法により，少数しか出現しない要素を特徴語として利用できる．

6 プロトタイプシステムの実装

6.1 システム構成

本稿で示したデータ作成及び提案手法に基づきキャラクタ検索のプロトタイプシステムを構築した．システムの実装にはフロントエンドとして Vue.js（ver.2.6.11）[12]，バックエンドとして DjangoRestFrameWork（ver.3.11.0）[13]，データベースとして MySQL（ver.8.0.21）[14]を利用した．処理手順を図 2 に示す．

6.2 インタフェース

本稿で作成した検索フォームは 2 つ存在し，1 つ目がキャラクタ検索であり 2 つ目が要素検索のフォームである．キャラクタ検索フォームに，ユーザの既知のキャラクタのフルネーム及び一部の名前を検索フォームに入力することにより，名前に該当したキャラクタが検索候補として出力される．検索候補として提示されたキャラクタをクリックすることによりキャラクタの持つ性格と外見の要素が可視化される（図 3 参照）．その中からユーザが加減算する要素を選択するために，要素をクリックすることで加減算を可能とさせた．要素をクリックすることで提示されている色が変更され，黄色が加算，青色が減算となる．

キャラクタの検索後サーバ側にキャラクタの演算式が送信される．サーバで要素同士の加減算が行われ，cos 類似度が高い上位 10 キャラクタが検索キャラクタとしてユーザに提示される．その際に，提示されるのは「性格と外見，性格のみ，外見のみ」のそれぞれで計算された上位 10 キャラクタである．タブを切り替えることにより，それぞれ算出されたキャラクタを確認することができる．検索されたそれぞれのキャラクタは詳細ボタンをクリックすることにより自由参加型 Web 百科事典から収集された文章とキャラクタに付与された要素を見ることができる．また，ピクシブのサイト[15]を利用することにより，任意のユーザが投稿したキャラクタのイラスト画像のページを閲覧可能となる．

7 実験

本稿では，キャラクタの構成要素の性格と外見に着目した加減算を行うことにより，類似キャラクタの検索が可能であるかを実験により確認する．実験では，上述のように構築したシステムをユーザに利用してもらうことによりキャラクタ検索の有効性を検証した．実験協力者は普段からコミックを読む学生を対象とした．

7.1 実験手順

実験協力者は，関西大学総合情報学部に在籍する学生の 20 名（男性 10 人，女性 10 人）であった．実験に先立ち，実験協力者に対して（1）実験の目的,（2）提案システムの操作説明,（3）実験課題の 3 つの項目を説明した．提案システムの操作説明では，キャラクタ検索（e.g., 入力フォームにキャラクタの一部の名前を入れると検索できる，キャラクタの名前をクリックすると要素が可視化され取捨選択できる）を説明した．また，検索されたキャラクタに対して，自由参加型 Web 百科事典のキャラクタ情報を閲覧できること，キャラクタをピクシブ百科事典のイラストページを用いることで，キャラクタのイラスト画像を確認できることを伝えた．これにより，全ての実験参加者が課題に用いるシステムの機能を理解している状態にした．

本実験の課題は「ユーザの既知のキャラクタを基に性格と外見の加減算に基づくキャラクタ検索の有効性の検証」である．この点について有効性を示すため提案システムを実験者に利用してもらった．実験者にはまず，検索キャラクタとしてあらかじめ用意されたコミックの主人公 15 人の中から，ユーザ既知のキャラクタを演算してもらい，次に，用意された以外のキャラクタのうち実験者の既知のキャラクタを演算してもらった．演算では「キャラクタ同士」と「キャラクタ＋要素」の 2 つの演

[12] https://jp.vuejs.org/index.html
[13] https://www.django-rest-framework.org/
[14] https://www.mysql.com/
[15] https://www.pixiv.net/

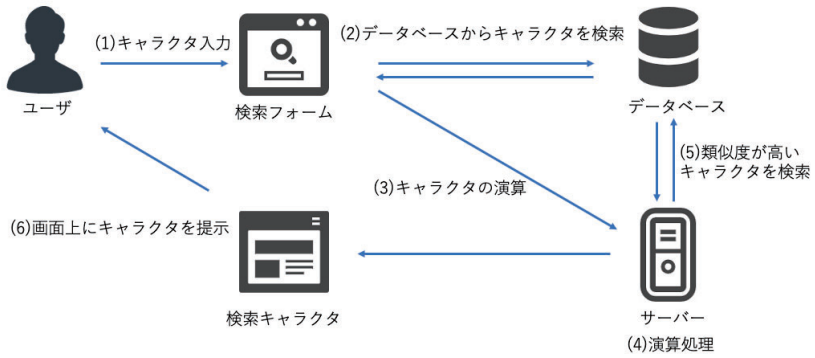


図 2 作成したプロトタイプシステムの処理手順

図 3 プロトタイプシステムの検索フォーム

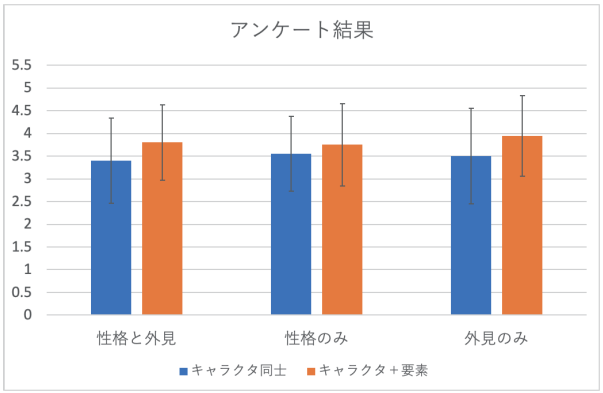


図 4 ユーザ実験によるアンケート結果

算を行うように指示した．実験中，演算の結果検索されたキャラクタに関して思ったことや感じたことなどについてあれば，自由にメモを記述するように促した．実験では，システムを利用する時間を 20 分経過した時点で終了とした．実験終了後，検索されたキャラクタに対して検索精度についてアンケートを回答してもらった．アンケートは「キャラクタ同士」の演算において演算したキャラクタの双方の特徴が反映されていたか,「キャラクタ＋要素」において演算した要素の特徴が反映されていたか，についてそれぞれ 5 段階評価で回答してもらった．アンケートには「キャラクタ同士」と「キャラクタ＋要素」の両者で「性格と外見，性格のみ，外見のみ」のそれぞれで回答項目を用意した．

7.2 実験結果

「キャラクタ同士」と「キャラクタ＋要素」の演算についてそれぞれのアンケート結果を図 4 に示す．グラフの結果はユーザの平均を表し，エラーバーは標準偏差である．表 4 にはアンケート結果の平均と標準偏差を示す．

7.3 考察

ユーザアンケート結果の評価値と自由回答からキャラクタの構成要素である性格と外見の加減算に基づくキャラクタ検索の有効性について考察する．

表 4 を確認すると,「キャラクタ＋要素」の結果が「キャラクタ同士」と比べて「性格と外見，性格のみ，外見のみ」の平均値を 0.2 ポイント程度上回り,「キャラクタ＋要素」の方が要素の特徴を反映した結果となった．

「キャラクタ＋要素」の検索結果が「キャラクタ同士」を上回った理由として実験者の自由回答を確認したところ,「性格に関しては少し片方の特徴が反映されたキャラが多く感じた」「どちらか一方のキャラクタに偏っている場合もあったと感じた」「適切なキャラが推薦されたと感じるがたまに一つのキャラクタの特徴に左右されている」という意見が見られた．「キャラクタ同士」の演算では両方のキャラクタの特徴ではなく片方のキャラクタの特徴が反映された結果に一部なっていたことが伺えた．これは,「キャラクタ同士」の演算を行った時，キャラクタの要素が同等の数キャラクタに付与されていないことが考えられる．また,「外見がぴったり当てはまっていても性格があまり演算されていないように感じた」「「帽子」などの装飾品の要素が強かった」という意見が得られ，出力結果が一部の要素に強く影響されることが原因である．

一方で,「キャラクタ＋要素」の検索結果では「キャ

表 4 ユーザアンケート結果の平均

演算方法	性格と外見	性格のみ	外見のみ
キャラクタ同士の演算	3.40 ± 0.94	3.55 ± 0.83	3.50 ± 1.05
キャラクタ＋要素の演算	3.80 ± 0.83	3.75 ± 0.91	3.95 ± 0.89

ラに外見の要素を足すとしっかり同じような性格のキャラで外見の違うキャラが出てきてとても精度が高いと感じた．性格の要素を足しても同じように感じた」「外見の特徴がそっくりな「優しい」要素を足したキャラクタを検索した時に想定したキャラクタが一番に出てきた」「選択したキャラクタをベースに要素を付け加えるので思った通りのキャラクタが出やすいように思った」というように出力された結果が想定したものになりやすい傾向があった．これらの結果から,「キャラクタ＋要素」によって，出力されたキャラクタを想定できることが示唆された．しかし，実験者数が少数であるため実験結果の平均値に大きな差が表れなかった．そのため今後は実験者数を増やすことを検討する．

8 まとめ

本稿では，キャラクタ情報に基づくコミック検索の実現を目指している．その端緒として，キャラクタを構成する要素の中でも性格と外見に着目したベクトル作成を試み，要素を加減算することにキャラクタを検索するプロトタイプシステムを提案した．ユーザ実験より,「キャラクタ＋要素」は出力されたキャラクタを想定できることが示唆され，性格と外見要素の加減算の有効性を確認できた．

今後の展望としてプロトタイプシステムの改善を検討する．本稿で行った実験の自由回答からは「外見でも性格でも抽出されたキャラクターが何の要素から導いたものなのかが分からなかった」「キャラクタのどの要素を重視しているかという個人差が大きい」という意見が得られた．構築したプロトタイプシステムではキャラクタに付与された要素を可視化するに留まり，キャラクタが出力される際にどの要素が重視されているかユーザに提示されていなかった．そのため出力されたキャラクタがどのような観点で出力されたかユーザは理解することが困難となる．以上の理由からも，今後はどの要素を重視してキャラクタが出力されたかを理解することができるシステムの構築を目指す．

参考文献

[1] 倉本到: 萌え擬人化キャラによるインタラクティブシステムの理解促進, 情報処理学会研究報告（EC）, Vol. 2012-EC-23, pp. 1–6 (2012).
[2] 松井俊樹, 朴炳宣, 松下光範: 萌え要素の加減算に基づくキャラクタの類似度判定手法の提案, 第 21 回インタラクティブ情報アクセスと可視化マイニング研究会, SIG-AM-21-01, pp. 1–6 (2019).
[3] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, *CoRR*, Vol. abs/1310.4546 (2013).
[4] 朴柄宣, 居林香奈枝, 松下光範: エゴグラムに基づいたコミックキャラクタの性格分類, 人工知能学会全国大会論文集, 1J3-02 (2018).
[5] 出版科学研究所: 出版月報 (2020 年 2 月号), 全国出版協会 (2020).
[6] 和田迪子, 渡部麻美, 市村美帆, 松井豊: Web 調査による新しいエゴグラムの尺度開発, 筑波大学心理学研究, Vol. 53, pp. 63–71 (2017).

ARG WI2 No.16, 2020
WI2-2020-23

コミックの登場人物についての説明文からの性格タグ推定

樋口 亮太　　山西 良典　　松下 光範
関西大学総合情報学部
{k896930, ryama, m_mat}@kansai-u.ac.jp

概要　本稿では，ストーリに基づいたコミックの検索を目指し，ストーリの構成要素のひとつであるキャラクタの性格に着目する．ストーリは,「どのようなキャラクタ」が「どのように活躍するか」で捉えられると考えた．このうち,「どのようなキャラクタ」に相当するキャラクタの性格などを示した文章は Web 上に存在するが，その記述は作品によって多様であるため，説明文自体をクエリとしてコミックの検索に用いることはできない．一方で,「どのようなキャラクタ」であるかを作品を超えて同一の基準で端的に表すために性格タグを用いることもある．しかし，性格タグの付与には膨大なコストと性格タグに対する解釈の曖昧性が存在する．本稿では，単語分散表現を用いて表現した Web 上のキャラクタ説明文と性格タグの関連性を機械学習によりモデル化し，説明文からの性格タグを推定する．推定結果の分析から，キャラクタの説明文の記述の傾向と性格タグとの関係性を考察する．

キーワード　コミック工学，ストーリ情報，キャラクタの性格，コンテンツ検索

1　はじめに

コミックの発行部数は毎年 1 万点以上にも及び，膨大である．この膨大なコミックの中からユーザが読みたい作品を選択する方法として，電子コミックや書籍販売のウェブサイトでは，コミックの検索サービスがある．現状の代表的なコミック検索法では，技術書や雑誌などの検索と同様に，書誌に関する情報（e.g., タイトル，作者，掲載誌）やジャンル（e.g., ラブコメ，アクション・アドベンチャー，ヒューマンサスペンス）などがクエリとして利用されている．しかし，コミックは，ストーリ性をもったコンテンツであり，これらのメタ的な情報だけではコンテンツの内容を表現できているとは言えない．現状のメタ情報のみを用いた検索の枠組みでは，ストーリ情報（例えば,「熱血なキャラクタが敵を倒すストーリ」や「優しいキャラクタが仲間を救うストーリ」）に基づいた検索は実現されていない．レビュー情報を用いてコミックの内容に基づいた検索 [3] も提案されているが，レビュー情報はコミックに対する読者の感想でありコンテンツの内容そのものを扱っているわけではない．また，レビュー情報にはネタバレが含まれる可能性 [1] もあり，コミックの内容そのものにフォーカスした検索手法が必要であると考えられる．

ストーリに基づいたコミック検索を実現するために，まずコミックのストーリー性を構成する要素について検討した．本研究では，“ストーリは「どのようなキャラクタ」が「どのように活躍するか」” で捉えられると考えた．本稿では，このうち「どのようなキャラクタ」に着目する．キャラクタを説明する上では，性格や生い立ちなどの内面的な情報と，見た目などの外見情報が用いられる．このうち性格は，キャラクタがどのように行動選択を行うかの大きな要因の一つとなり，ストーリ展開に大きく影響する可能性が高いと考えられる．性格に着目した研究の一つとして，朴らは性格診断に用いられるの一つであるエゴグラムの 5 つの性格的特徴を用いて，各性格特徴に対応した単語群を用意した辞書を作成し，キャラクタを説明する文章（以下，キャラクタの説明文）から性格分類を行った [2]．しかしながら，あらかじめ著者らが用意した辞書を用いたアプローチであるため，辞書に含まれない多用な表現で記述されたキャラクタの説明文に対しての性格分類の有効性においては疑問が残る．また，性格もエゴグラムの 5 つの性格特徴を用いたベクトル表現としているが，様々なコミックのキャラクタの特徴を表現するうえで，5 次元ベクトルで十分であるかについても検討しなければならない．ファンコミュニティの中では，キャラクタの性格を端的に表すために，性格タグが用いられることもある．タグでの表現により，異なる作品に登場するキャラクタの性格を多様性を担保しつつも，多くの人の中で共通の理解で評価可能になる．そこで，本稿では Web 上に複数存在するキャラクタを説明文そのものを入力として，ファンコミュニティの中で共通了解が得られていると考えられる性格タグを推定する．

提案手法では，単語分散表現を用いて表現した Web 上のキャラクタ説明文からコミックキャラクタの性格タグを推定する．キャラクタの説明文と性格タグの関係性をモデル化できれば，未知のキャラクタに対しても説明文からタグを自動的に付与することが可能になる．これにより，キャラクタの性格をコミックの検索に導入し，ストーリに基づくコミック検索の端緒となることを目指す．

Copyright is held by the author(s).
The article has been published without reviewing.

表 1　キャラクタの説明文についての情報抽出源とそこから得られた説明文数.

自由参加型ウェブ百科事典	説明文数
Wikipedia[1]	535
ニコニコ大百科[2]	441
pixiv 百科辞典[3]	538
アニヲタ wiki（仮）[4]	490
合計	2,004

表 2　キャラクタに付与された頻出上位 20 件の性格タグと対応する説明文数.

性格タグ	説明文章の数
強気	1,059
異性が好き	1,038
優しい	905
俺	831
色恋に興味はある様子	742
気の強さは普通	685
プライド高い	670
精神年齢が高い	604
堂々とした	598
精神年齢が若い	562
まとも	530
ちょっと変	514
私・わたし	509
社交的	486
普通の上品さ	473
S	423
世話焼き	416
お調子もの	370
落ち着いた	356
紳士的	344

2　データセット

Web 上からキャラクタの説明文と性格タグを自由参加型ウェブ百科事典からそれぞれ収集し，データセットを構築した．キャラクタの説明文については，複数の自由参加型ウェブ百科事典から横断的に収集した．

キャラクタの説明文については，表 1 に示すように，複数の自由参加型 Web 百科事典を情報リソースとして，コミックのキャラクタについて記述している説明文を取得した．得られたキャラクタの種類数は 537 人であり，複数の情報リソースに記述されているキャラクタもいれば，いずれかの情報リソースにしか記述されていないキャラクタもいた．同じキャラクタについて説明していても，情報リソースによって説明文の記述には大きな差異が見られる．例えば,「ジョジョの奇妙な冒険」における「空条承太郎」について Wikipedia では設定に関する客観的な情報に関する記述が多い一方で，アニヲタ wiki（仮）では,

> 『道とは自分で切り開くもの』などその行動の基盤には確固たる信念があり、いかなる詭弁や恫喝を向こうに回しても一切のブレが無い。

のように，キャラクタのセリフが引用されて説明されている．そこで，本稿では同一のキャラクターを説明していたとしても異なる情報リソースから得られた説明文については，それぞれ異なる記述特徴を有した独立の説明文として扱うこととした．

キャラクタの性格タグについては，キャラ属性王国[5]を情報リソースとして収集した．キャラ属性王国では，コミックのキャラクタを説明するために,「ステータス」「容姿」などの外見情報や,「背景」「能力」などの内面的な情報をカテゴリとして様々なタグが付与されている．本稿では，このうちキャラクタの性格に関連する「基礎性格」「性格」のカテゴリに分類されているタグを収集した．上述の説明文が取得できたキャラクタの性格を表現するためには，合計で 225 種類のタグが用いられていた．本稿では，得られた性格タグのうち，表 2 に示す頻出上位 20 件の性格タグを対象とした．1 人のキャラクタには複数の性格タグが付与されているため，各説明文は 20 種類の性格タグのそれぞれが付与されたかバイナリで表現された 20 次元の性格ベクトルを持つことになる．

3　提案手法

提案手法では，以下の手順でキャラクタの説明文から性格タグの推定を行う．

1. 単語分散表現によって，説明文を多次元ベクトル化する
2. 入力を上記 1 で得られた多次元ベクトル，出力を性格タグを示す 20 次元のベクトルとしてキャラクタの説明文と性格タグの関係モデルを学習する．
3. 学習したモデルを用いて，キャラクタの説明文から性格タグを示すベクトルを推定する．

下節では，手順 1 および 2 の詳細について説明する．

3.1　単語分散表現の獲得

説明文の多次元ベクトル化に用いる単語分散表現を獲得するために，まず 2 章で用意した説明文 2,004 件を学習した．ここで，学習した全 2,004 件の説明文には，42,412 文，1,051,036 単語が含まれていた．学習に際して，説明文を形態素解析によって単語単位に分割し，その表層形を学習した．形態素解析器には MaCab（ver.0.996），辞書には NEologd（ver.0.0.7）[6]を用いた．辞書 NEologd には，有名なコミックのキャラクタ名も含まれている．

[5] https://chara-zokusei.jp
[6] https://github.com/neologd/mecab-ipadic-neologd

表 3 各性格タグの推定における 5 分割交差検証での平均評価：Precision，Recall，F1 値.

性格タグ	平均 Precision	平均 Recall	平均 F1 値
強気	0.699	0.724	0.699
異性が好き	0.718	0.712	0.718
優しい	0.658	0.621	0.658
俺	0.674	0.678	0.674
色恋に興味はある様子	0.640	0.654	0.640
気の強さは普通	0.577	0.603	0.577
プライド高い	0.575	0.664	0.575
精神年齢が高い	0.575	0.608	0.576
堂々とした	0.508	0.520	0.509
精神年齢が若い	0.488	0.597	0.488
まとも	0.518	0.606	0.519
ちょっと変	0.454	0.502	0.454
私・わたし	0.557	0.606	0.557
社交的	0.456	0.545	0.457
普通の上品さ	0.444	0.535	0.444
S	0.482	0.548	0.483
世話焼き	0.475	0.562	0.476
お調子もの	0.417	0.532	0.385
落ち着いた	0.417	0.532	0.418
紳士的	0.425	0.520	0.426

単語分散表現の学習では，ライブラリ gensim の W2V (ver.3.8.3)[7]を利用した．このとき，単語分散表現の学習時の window サイズは 5，ベクトルの次元数は 100 とした.

3.2 説明文と性格タグの関係モデルの学習

キャラクタの説明文と性格タグの関係モデルを学習するためには，3 層のニューラルネットワーク（NN）を用いた．NN の学習には，ライブラリ PyTorch[8]を用いた.

ネットワークの構成は，第 0 層（入力層）：100，第 1 層（隠れ層 1）：500，第 2 層（隠れ層 2）：250，第 3 層（出力層）：20 とした．入力層には，3.1 節で得られた 100 次元の単語分散表現ベクトル，出力層には，20 次元の性格タグベクトルがそれぞれ対応する．推定においては，各性格タグ毎に推定モデルを構築する方法も考えられるが，推定モデルの構成については今後の検討課題とする.

4 実験

提案手法によって，キャラクタの説明文の記述から性格タグを推定可能であるか実験した．下節にて，実験設定と実験結果，考察を述べる.

4.1 実験設定

実験では，説明文 2,004 件を対象として，5 分割交差検証を行った．このとき，交差検証ごとに NN の学習モデルを構築し直し，テストデータには学習に用いていない説明文を用いた.

学習した NN では，説明文を入力すると，その説明文に対応する性格タグの 20 次元のベクトルを出力する．実験の評価では，性格タグそれぞれについての Precision, Recall，マイクロ F-1 値を算出した．本稿では，これらの評価指標の 5 分割検証実験における平均値によって推定性能について議論することとする.

4.2 結果

表 3 に，推定結果を示す．F1 値を見てみると,「強気」「異性が好き」「優しい」「俺」などのタグについては比較的高い平均 F1 値が得られた．一方で,「お調子もの」「落ち着いた」「紳士的」などのタグについては，比較的低い F1 値の推定結果となった.

これらの結果から，高い推定性能を示すとまでは言えないものの，多様な性格タグをキャラクタの説明文を入力として用いることで推定できる可能性が示唆された．一方で，分散表現の獲得や関係モデルを構築するネットワークの構成や学習方法などについては，さらなる検討が必要であると考えられる.

4.3 考察

表 3 の結果をもとに，キャラクタの説明文の記述について考察した．本稿で推定対象とした性格タグでは，対義関係にあるタグ「精神年齢が高い」と「精神年齢が若い」のペアや「お調子もの」と「落ち着いた」のペアが存在したり，どちらも第一人称である「俺」と「私・わたし」などがあった．また,「ちょっと変」のような一見して性格の詳細とは紐つかないタグも存在した.

「俺」と「私・わたし」のペアはサンプル数の違いを考慮しても,「俺」と「私・わたし」の平均 F1 値の差が大きい．例えば,「俺」という性格タグが付与されているキャラクタには「鬼滅の刃」に「嘴平伊之助」が登場

[7] https://radimrehurek.com/gensim/
[8] https://pytorch.org/

する．このキャラクタの Wikipedia には，「戦う相手が居ない場でもその闘争心が収まらないらしく、大声を張り上げたり意味も無く木の幹に体当たりしたりしている。」という強気の男性とわかる記述がある．一方で，「私・わたし」という性格タグが付与されているキャラクタには女性にも多いが，中性的な男性にも付与されていることがある．例えば,「私・わたし」という性格タグは「七つの大罪」に登場する「エリザベス」に対して，「HUNTER × HUNTER」に登場する「クラピカ」にも付与されている.「俺」の性格タグが付与されているキャラクタは比較的強気の男性キャラクタに限定されるが，「私・わたし」の性格タグが付与されているキャラクタは，性別が混在しており，説明文の内容が限定されにくいため，推定結果に差が出たと考えられる．

「ちょっと変」という性格タグは全体的に低い推定結果になっている．この性格タグが付与されているキャラクタ説明文には，他にも様々な性格タグが付与されている場合が多く，内容的にもまとまりがないことが低い原因だと考えられる．例えば,「ちょっと変」という性格タグが付与されているキャラクタの「新世紀エヴァンゲリオン」に登場する「碇シンジ」の pixiv 百科事典には「やや内省的で繊細な性格。自らの存在意義に思い悩んでおり、苛酷な状況に追い詰められた際などは極めて情緒不安定に陥る事も。」という記述がある．また，同じ作品に登場する「葛城ミサト」にも「ちょっと変」という性格タグが付与されている.「葛城ミサト」の Wikipedia には「私生活においては、非常にがさつかつ、ずぼらでだらしない面が多い。」と記述されている．同じ作品の同じ性格タグが付与されているキャラクタの説明文であっても，その説明の記述は大きく異なる．このような性格タグは説明文からの推定が難しいことがわかった．

5　おわりに

本稿では，キャラクタの説明文とキャラクタの性格タグをウェブ上の異なる複数の情報リソースから収集し，それらを組み合わせることによってコミックのコンテンツであるキャラクタの性格を推定した．キャラクタの説明文とキャラクタの性格タグの関係モデルの構築では，単語分散表現を用いて表現した Web 上のキャラクタ説明文と性格タグの関係性を獲得した．実験の結果，多様な性格タグに対して，一定の推定性能で説明文から性格タグが推定できる可能性が示唆された．

提案手法による推定性能について考えると，推定器のネットワークの構成やパラメータの調整など検討すべき点は多い．これらについては，今後の課題とする．また，性格がストーリにどのような影響を与えるのかについて調査し，コミックのストーリ検索の実現を目指す．

謝辞

本研究は，一部，文科省科研費基盤 C#20K12130 の助成のもと行われた．本稿の執筆にあたって，関西大学総合情報学部小林達哉の協力を得た．記して謝意を表す．

参考文献

[1] 佐藤剣太, 牧良樹, 中村聡史: 未読および既読シーンの提示が読者のコミック閲覧意欲に与える影響, 情報処理学会研究報告 (EC), Vol. 2018-EC-47, No. 3, pp. 1–8 (2018).
[2] 朴柄宣, 居林香奈枝, 松下光範: エゴグラムに基づいたコミックキャラクタの性格分類, 人工知能学会全国大会論文集, 1J3-02 (2018).
[3] 山下諒, 朴炳宣, 松下光範: コミックの内容情報に基づいた探索的な情報アクセスの支援, 人工知能学会論文誌, Vol. 32, No. 1, pp. WII–D_1–11 (2017).

ARG WI2 No.16, 2020
WI2-2020-24

Web 人名検索結果の要約と可視化を目指して
－2010 年代の進捗－

村上 晴美

大阪市立大学大学院工学研究科

harumi@osaka-cu.ac.jp

概要　著者の研究室では，テキストやデータからの人物の理解に関する研究を行っている．その中で，Web 上の人物の理解のために，Web 人物検索における要約と可視化の研究を行っている．研究の目的は，Web 上の人名検索においてユーザによる人物の選択と理解を支援するインタフェースの開発である．先行研究において 2000 年代の研究の進捗報告を行った．本稿では，2010 年代の研究の進捗の概要をまとめる．内容は，NDC 人物ディレクトリ，人間による同姓同名人物の分離，履歴書と地図の表示，Wikipedia 風概要文の作成，件名の付与である．

キーワード　Web 人物検索，進捗報告，要約

1 はじめに

著者の研究室では，テキストやデータからの人物の理解に関する研究を行っている．その中で，Web 上の人物の理解のために，Web 人物検索における要約と可視化の研究を行っている．先行研究[1, 2]において 2000 年代の研究の進捗報告を行った．本稿では，2010 年代の研究の進捗の概要をまとめる．

2 研究の概要

研究の目的は，Web 上の人名検索においてユーザによる人物の選択と理解を支援するインタフェースの開発である．

Web 人名検索結果を同姓同名人物毎に分離し，人物クラスタ（人物毎の HTML 文書群）を作成する．ユーザによる人物の選択と理解を支援するために，検索結果と人物の要約と可視化を行う．

本研究では「抽出した情報を用いて人物の選択や理解に有用な情報を作成すること」を「要約」と呼ぶ．要約には抽出した情報の統合や，他の情報源の情報の取得を含む．

3 先行研究の概要

2000 年代の主な内容は，初期プロトタイプ，同姓同名人物の自動分離，職業関連情報の抽出，位置情報の取得[1, 2]及び，履歴書の作成[3]であった．この中で，職業関連情報の抽出と履歴書の作成については研究を完了した．

以下では，先行研究以降の進捗を述べる．

Copyright is held by the author(s).
The article has been published without reviewing.

4 NDC 人物ディレクトリ

人物に図書館の分類体系である日本十進分類法（NDC）の分類記号を付与する手法を提案した．人物に図書館の分類記号を付与することにより，人物にラベルを付与すると同時に図書館の分類体系を利用した人物ディレクトリを構築できる．関連する語（索引語）と分類記号を関連付ける NDC の相関索引を利用し，タイトル要素のテキストに含まれる索引語の頻度に基づき分類記号を求める．人物クラスタに対して上位 5 件の NDC9 を付与して人物ディレクトリを試作した．20 氏名×100 件の Web 検索結果から得た 152 人物クラスタを対象として，4 手法×6 文書を組み合わせた評価実験を行った．相関索引とタイトル要素のテキストを用いて人物に NDC の分類記号を付与する手法の有効性を確認した．

本研究は文献[4-8]において発表した．

5 人間による同姓同名人物の分離

人がどのように Web 上の同姓同名人物を判別するか調査を行った．20 氏名×20 件の Web 検索結果を分けるように教示し，質問紙，プロトコル分析，インタビューを用いて判別プロセスを分析した．キーワード，職業，作品（架空の人物の場合は登場する作品），顔画像，関連する人名が，人物を識別するために重要であることがわかった．実験結果に基づき同姓同名人物の分離モデルと知識構造モデルを提案した．

本研究は文献[9, 10]において発表した．

6 履歴書と地図の表示

人物の履歴書と地図を表示するシステムを試作した．

提案手法は人物クラスタに対する (1) 先行研究[3]を用いた履歴書の作成, (2) 履歴書中の学歴と経歴から学校と勤務先の抽出, (3) 学校と勤務先から位置情報の取得, (4) 履歴書と地図の表示の 4 段階からなる. 56 氏名(主として有名人)×50 件の Web 検索結果から作成した 56 人物クラスタを対象として, 学校と勤務先の抽出及び位置情報取得の評価実験を行った.

本研究は文献[11-13]において発表した.

7 Wikipedia 風概要文の作成

人物の概要文を作成する手法を提案した. 人物クラスタから属性情報(氏名のよみ, 生年月日, 没年月日, 出身地, 職業, 所属, 役職)を抽出し, Wikipedia の第一文風の概要文を作成する. 20 氏名×50 件の Web 検索結果から得た 80 人物クラスタを対象として評価実験を行った. Wikipedia 風テンプレートと, 抽出された属性情報の組合せにより Wikipedia の第一文風の概要文を作成できることがわかった.

本研究は文献[14, 15]において発表した.

8 件名の付与

国立国会図書館の提供する件名標目表である NDLSH を人物に付与する手法を調査した. NDLSH を人物に付与することにより, ゴミの少ないキーワードを付与すると同時に関連語を用いた探索的な検索が可能となる. (a)検索ランキング, (b) HTML 文書内の位置, (c) 同義語の使用, (d)文書頻度の使用の組合せについて, 7 節と同様にして得た 80 人物クラスタを対象とした 405 (5×9×3×3) パターンの実験を行った. 実験環境においては, (a)上位 10 件, (b) 氏名の前後 100 文字(i.e., 200 文字), (c) 同義語の重み 0.5 倍, (d) 文書頻度の使用の組合せがよかった.

また, 日本図書館協会が提供する件名標目表である BSH4 においても同様の実験を行った. (a) 上位 10 件と (c) 文書頻度の使用が良いことが共通した.

本研究は文献[16-18]において発表した.

9 おわりに

NDC 人物ディレクトリについては研究を完了した. 完了していない研究については 2020 年代の課題として研究を続けている. 本稿作成時点においては, 概要文の作成と件名の付与の研究を行うと同時に, ネットワークインタフェースの試作とキーワード抽出の研究を行っている.

謝辞

本研究は JSPS 科研費 19K12718 の助成を受けたものです.

参考文献

[1] 村上晴美, 上田洋:Web 人名検索結果の要約と可視化を目指して, 2009 年度人工知能学会全国大会(第 23 回)論文集, 2009.

[2] Murakami, H., Ueda, H., Kataoka, S., Takamori, Y. and Tatsumi, S.: Summarizing and visualizing Web people search results, ICAART 2010, Vol. 1, pp. 640-643, 2010.

[3] 上田洋, 村上晴美, 辰巳昭治:Web 上の人物理解のための履歴書作成, 人工知能学会論文誌, Vol. 25, No. 1, pp. 144-156, 2010.

[4] 浦芳伸, 村上晴美, NDC を用いた人物ディレクトリの開発, 情報処理学会第 73 回全国大会講演論文集, Vol.1, pp. 651-652, 2011.

[5] Murakami, H. and Ura, Y.: People search using NDC classification system, ESAIR 2011, pp. 13-14, 2011.

[6] Murakami, H., Ura, Y. and Kataoka, Y.: Assigning library classification numbers to people on the Web, AIRS 2013, LNCS, vol 8281. Springer, pp. 464-475, 2013.

[7] 片岡祐輔, 浦芳伸, 村上晴美:NDC を用いた人物ディレクトリの評価実験, 電子情報通信学会 2013 年総合大会情報・システムソサイエティ特別企画学生ポスターセッション予稿集, pp. 34, 2013.

[8] 村上晴美, 浦芳伸, 片岡祐輔:Web 上の人物への図書館の分類記号の付与と人物ディレクトリの開発, システム制御情報学会論文誌, Vol. 29, No. 2, pp. 51-64, 2016.

[9] 三宅悠生, 村上晴美:人は Web 上の同姓同名人物をどのように判別しているのか, 電子情報通信学会第二種研究会資料(第 19 回 Web インテリジェンスとインタラクション研究会), pp. 73-76, 2011.

[10] Murakami, H. and Miyake, Y.: How do humans distinguish different people with identical names on the Web?: a cognitive science approach, CIKM 2012, pp. 2475-2478, 2012.

[11] 王爽, 浦芳伸, 上田洋, 村上晴美:Web 上の人物履歴情報の地図上への表示, 電子情報通信学会 2012 年総合大会情報・システムソサイエティ特別企画 学生ポスターセッション予稿集, pp. 99, 2012.

[12] 唐春亮, 王爽, 上田洋, 村上晴美:Web 上の人物履歴情報の地図表示システム, 2013 年度人工知能学会全国大会(第 27 回), 2013.

[13] Murakami, H., Tang, C., Wang, S. and Ueda, H.: Vitae and map display system for people on the Web, IEA/AIE 2014, LNCS vol 8482. Springer, pp. 348-359, 2014.

[14] 村上晴美, 小西利宗, 浦芳伸:Web 上の人物の概要文の作成, 2016 年度人工知能学会全国大会(第 30 回), 2016.

[15] Murakami, H., Konishi, T. and Ura, Y.: Generating Wikipedia-like biographical sentences from Web people search results, IIAI-AAI 2017, pp. 992-993, 2017.

[16] 下倉雅行, 村上晴美:Web 上の人物への NDLSH の付与, 2017 年度人工知能学会全国大会(第 31 回), 2017.

[17] Shimokura, M. and Murakami, H.: Assigning NDLSH headings to people on the Web, AIRS 2017, LNCS, vol 11292. Springer, pp. 189-195, 2018.

[18] 下倉雅行, 村上晴美:Web 上の人物への BSH の付与, 2018 年度人工知能学会全国大会(第 32 回), 2018.

スポンサー

シルバー

事務局:アカデミック・リソース・ガイド株式会社

WI2 研究会事務局

〒231-0012　神奈川県横浜市中区相生町 3-61　泰生ビル
さくら WORKS＜関内＞407