ARG WI2 No.10, 2017
WI2-2017-3

文重要度と図表引用文の位置情報を用いた図表の重要度推定

平岡 誉史† 山西 良典†† 福本 淳一†† 西原 陽子††

† 立命館大学大学院情報理工学研究科 †† 立命館大学情報理工学部

{*is0230ff@ed, ryama@media, fukumoto@media, nisihara@fc*}*.ritsumei.ac.jp*

概要 本稿では，複数のメディアが組み合わされたマルチメディア情報において，メディア間で重要度を伝搬させることで重要度推定が困難なメディアに対して重要度を推定する手法について述べる．本稿で処理対象として扱う科学技術論文は，言語情報である本文と図表の画像情報を組み合わせることで，複雑な内容を読者が容易にまた詳細に理解可能としている．提案手法では，論文中での単語の出現頻度をもとに推定した各文の重要度を，図表引用文との位置情報を用いて図表へ伝搬することで重要度を推定する．本稿では，口頭発表された論文中でポスター発表時のポスターに採用された図表を，重要度が高い図表と仮定した重要な図表の抽出実験を行った．提案手法と２つの比較手法それぞれを用いて重要度推定を行い，推定結果から算出した平均適合率を比較することで提案手法の有効性を評価した．評価実験の結果，提案手法は最も高い平均適合率を示し，精度の高い図表の重要度推定が行えることを確認した．

キーワード 重要度伝搬，図表の重要度推定，創作支援，ポスター作成

1 はじめに

現在の我々の身の回りの情報には，複数の異なるメディアの情報を組み合わせて用いたマルチメディア情報が多く存在する．例えば，音楽は音情報と言語情報を組み合わせたもの，動画は音情報と画像情報を組み合わせたものと考えられる．学術的な分野では論文もその１つと言える．本稿では，複数の異なるメディアが組み合わされたマルチメディア情報において，メディア間で重要度を伝搬させることで，重要度の推定が困難なメディアに対しての重要度推定を実現するための手法を提案する．

研究者は研究を遂行するにあたって先行研究を含めて自身の研究に関連した数多くの論文に目を通す必要があるが，論文の多くは専門性が高く，論文の閲読と内容の理解には，膨大な時間を費やしてしまう．論文には，詳細な内容の把握やデータ確認を容易にするための図表が含まれている．また，論文を基にした口頭での研究発表も行われる．このとき，論文を基にして作成されるポスターや発表スライドは論文の内容を簡潔にまとめたものであるが，多くのポスターやスライドにおいて論文中で使用された図表が利活用されている．

論文には数多くの図表が使われており，各図表には，その論文を理解する上での重要度が存在すると考えられる．論文読者は，重要だと思う文章箇所を重点的に読んで内容を理解するように，複数ある図表に対してもその中から特に重要な図表に注目して内容の理解を図る．また，論文からポスターやスライドを作成する際には，図１に示すように論文中のすべての図表を掲載するのではなく重要な図表を選択して掲載していることが多い．現在は論文の書き手や読み手が重要な図表の判別を行っているが，この図表の重要度判別を自動で行うことが可能になれば，書き手であればポスターやスライド作成時における一支援として，読み手であればより内容把握を容易にする支援として活かせるのではないかと考えた．そこで，本稿では論文中の図表の重要度を推定する手法を提案する．

2 関連研究

本稿では，言語情報と画像情報によるマルチメディア情報の要約を行う．動画要約に関する様々な研究では，画像情報の重要度推定手法がいくつか報告されている．動画要約は，動画から代表的な画像（以下，キーフレーム）を複数枚抽出して表示することで動画の要約を行う．動画要約では動画中の重要な部分を簡潔にまとめて表示するために，キーフレームの選択手法が重要になる．三浦ら [1] は，映像中の特徴的な動きを検出し料理映像中の重要なシーンをキーフレームとして自動で抽出している．Yao Ting ら [2] は，機械学習の技術を用いて各フレームの重要度推定をし，キーフレームの選択を行っている．また，笠松ら [3] は，画像内の人物を検出し，その大小や有無によって重要度推定を行う画像特徴を用いた手法を提案している．これらの動きなど動画特有の特徴を使って重要度推定を行う手法は，論文中の図表のような静止画に対しての重要度推定には向かない．画像特徴を用いた重要度推定は，論文中の図に対しては有効である可能性が考えられるが，論文中で図と同様に扱われる数値などのデータを掲載した表に関しては有効ではないと考えられ，図表の区別なく重要度推定を行うための手法を提案する必要がある．

論文中の図表に関しては，図表に関連した情報の抽出を試みる研究が行われている．特に図表説明文の抽出に

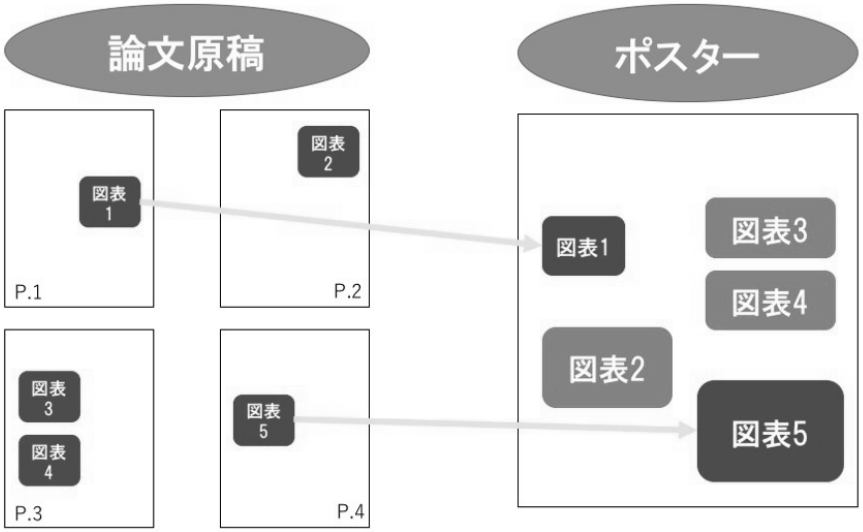

図 1 論文からポスターへの図表の引用の例.

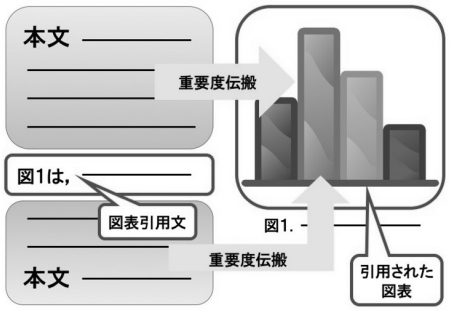

図 2　提案手法における文から図表への重要度伝搬のイメージ図.

関する研究として，市野ら [4] や竹島ら [5] の研究がある．これらの研究では,「図表を明示的に引用している文（以下，図表引用文）の前後数文は関連している文である可能性が高いこと」や「図表と関連している文は図表領域内の図表番号や単語が頻繁に出現すること」などの図表と文との関連特徴が挙げられている．言語情報の重要度推定に関しては，文章の自動要約の分野を中心として様々な研究が行われている．Zechner [6] の単語の重要度の和を文の重要度とする手法や川辺ら [7] のように機械学習手法を用いて重要度を求める手法など様々な重要度推定の手法が提案されている．本稿の提案手法では，論文が言語と図表（画像）から構成される点に着目し，これらの関連特徴を用いて文重要度を図表に伝搬することで図表の重要度推定を行う．

3　提案手法

本稿の提案手法は，図 2 に示したイメージ図のように，論文中の文重要度を図表に伝搬することで図表の重要度推定を行う．文重要度は論文中での単語の出現頻度と各文間の類似度をもとに算出する．文重要度の図表への重要度伝搬には，1 章で述べた市野ら [4] や竹島ら [5] の研究で挙げられていた図表と文との関連特徴を応用して，論文中での各図表の図表引用文との位置情報を用いる．

提案手法は，図表引用文の近くに重要度が高い文が存在すると，引用されている図表の重要度を高くする．また，論文中で複数回引用されている図表は，論文中に図表引用文が複数出現するので，各引用文の位置情報を用いて重要度を推定し，その総和を図表の重要度する．以下で，提案手法の詳細について述べる．

3.1　文重要度の算出

各文の初期重要度を単語の出現頻度を用いて算出する．各文からの単語抽出には MeCab [8] による形態素解析を用いた．重要度計算には「名詞」のみを使用し，式 (1) に従って文 i に対する初期重要度 $Timp(i)$ を求める．

$$Timp(i) = \sum_{w \in \boldsymbol{W(i)}} n(w,i) \times f(w). \qquad (1)$$

ここで，$n(w,i)$ は文 i 中での名詞 w の出現頻度，$f(w)$ は論文全体における名詞 w の出現頻度をそれぞれ示す．また，$\boldsymbol{W(i)}$ は文 i を構成する名詞集合を示す．

続いて，文同士の類似度を用いて文重要度を更新する．この文重要度の更新により，単語の出現頻度のみでは捉

えきれない，図表の説明文をさらに説明する文といった段階的な論理展開への対応を図る．論文中の文 i と j の類似度 $R(i,j)$ は cos 類似度を求める式 (2) に従って求められる．

$$R(i,j) = \frac{\sum_{w \in \boldsymbol{W(i)}, \boldsymbol{W(j)}} n(w,i) \times n(w,j)}{\sqrt{\sum_{w \in \boldsymbol{W(i)}} n(w,i)^2} \times \sqrt{\sum_{w \in \boldsymbol{W(j)}} n(w,j)^2}}, \tag{2}$$

ここで，$n(w,i)$ と $n(w,j)$ はそれぞれ i 番目の文と j 番目の文に出現する名詞 w の出現頻度を示す．最終的な i 番目の文重要度 $TIMP(i)$ は，得られた $R(i,j)$ を用いることで，式 (2) により得られる．

$$TIMP(i) = Timp(i) + \sum_{j \in \boldsymbol{N}} R(i,j) \times Timp(j). \tag{3}$$

ここで，$Timp(i)$ と $Timp(j)$ はそれぞれ式 (1) で求めた i 番目と j 番目の文の初期重要度，$R(i,j)$ は式 (2) で求めた i 番目と j 番目の文の類似度を示す．$\boldsymbol{N}$ は，論文を構成する文集合を示す．

3.2 図表引用文の位置情報を用いた図表への文重要度の伝搬

論文中から各図表を直接引用している文を検索し，引用文に近い文ほど大きく，遠い文ほど小さくなる重み付けを行う．図表 k における i 番目の文の重み $Cimp_k(i)$ は，式 (4) に従って算出される．

$$Cimp_k(i) = \sum_{r \in \boldsymbol{F_k}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{(i-r)^2}{2}), \tag{4}$$

ここで，r は図表の引用文番号を示し，$\boldsymbol{F_k}$ は図表 k を引用する文の集合を示す．式 (4) は，平均を r，標準偏差を 1 とした正規分布の式とみなせる．

式 (3) によって算出された文重要度と式 (4) で算出される図表引用文から算出される重みを用いることで，文重要度を図表に伝搬する．文重要度を図表 k へ式 (5) に従って伝搬することで，図表の重要度 $CIMP(k)$ を得る．

$$CIMP(k) = \sum_{i \in \boldsymbol{N}} Cimp_k(i) \times TIMP(i). \tag{5}$$

このとき，$CIMP$ の値域は論文によって異なるため，式 (6) に従って各論文中の図表の相対的な重要度 $CIMP'(k)$ に正規化する．

$$CIMP'(k) = \frac{CIMP(k)}{\sum_{k \in \boldsymbol{K}} CIMP(k)}, \tag{6}$$

ここで，$\boldsymbol{K}$ は論文中に出現する図表の集合を示す．

4 比較手法

提案手法の図表の重要度推定に関する有効性を確認するため，比較手法を 2 種類用意し，その精度の比較検討を行った．以下で，各比較手法の詳細について述べる．

4.1 比較手法 1：キャプション文と本文中の各文との類似度を用いた文重要度伝搬手法

比較手法 1 では，図表引用文との位置情報を用いずに文重要度を伝搬する．比較手法 1 と提案手法の比較からは，文重要度を伝搬する際の図表引用文との位置情報の有効性を確認する．比較手法 1 では，文重要度を図表に伝搬する際に，図表引用文と各文の位置情報に代わって各文と各図表のキャプション文との類似度を用いる．これは，市野ら [4] や竹島ら [5] の研究で挙げられている「図表と関連した文は図表領域内の図表番号や単語が頻出しやすい」という特徴を解釈したモデルと言える．以下で，比較手法 1 の詳細について述べる．

文重要度に関しては式 (1)〜(3) を用いて提案手法と同様に算出する．各図表のキャプション文 c_k と論文中の各文 i との類似度 $CR(c_k,i)$ は，式 (7) に従って算出する．

$$CR(c_k,i) = \frac{\sum_{w \in \boldsymbol{W(c_k)}, \boldsymbol{W(i)}} n(w,c_k) \times n(w,i)}{\sqrt{\sum_{w \in \boldsymbol{W(c_k)}} n(w,c_k)^2} \times \sqrt{\sum_{w \in \boldsymbol{W(i)}} n(w,i)^2}}, \tag{7}$$

ここで，$n(w,c_k)$ と $n(w,i)$ はそれぞれ図表 k のキャプション文と論文中の i 番目の文に出現する名詞 w の出現頻度を示す．$\boldsymbol{W(c_k)}$ は，キャプション文 c_k を構成する名詞集合を示す．

比較手法 1 における図表 k の重要度 $cCIMP(k)$ は，式 (8) で算出する．

$$cCIMP(k) = \sum_{i \in \boldsymbol{N}} CR(c_k,i) \times TIMP(i). \tag{8}$$

また，提案手法と同様に，式 (9) に従って論文中の図表における相対的な重要度に正規化して扱う．

$$cCIMP'(k) = \frac{cCIMP(k)}{\sum_{k \in \boldsymbol{K}} cCIMP(k)}. \tag{9}$$

4.2 比較手法 2：論文中での図表の面積を用いた手法

比較手法 2 と提案手法を比較することで，文重要度を図表に伝搬することの有効性を確認する．比較手法 2 で

は，文重要度を図表に伝搬せず，論文中での面積が大きい図表は重要度が高いと仮定し，論文中での図表の面積を図表重要度とする手法を用いる．論文の PDF データ中における図表画像領域を示す矩形の縦の長さを h，横の長さを w として，図表 k の面積 $S(k)$ を式 (10) で算出する．

$$S(k) = h(k) \times w(k). \tag{10}$$

5 評価実験

5.1 実験環境

論文中の重要な図表について，ポスター発表時に論文著者が用いたものを重要度が高いと仮定した．本稿では，第３０回人工知能学会全国大会 JSAI2016[1] にて，発表された論文とポスターを参照して評価実験を行った．

提案手法と 4 章で述べた 2 種類の比較手法の各手法で図表の重要度を推定し，推定結果から平均適合率を算出し，比較することで評価を行った．論文 p に対しての平均適合率 $AP(p)$ は，式 (11) に従って算出される．

$$AP(p) = \frac{1}{\boldsymbol{A_p}} \times \sum_{\boldsymbol{A_p}} Precision_a, \tag{11}$$

ここで，a はポスター中で用いられた図表インデックス，$Precision_a$ は図表 a の適合率をそれぞれ示す．また，$\boldsymbol{A_p}$ はポスター中で用いられた図表の集合を示す．

5.2 実験結果

表 1 に，実験対象の論文毎に算出した各手法を用いた図表重要度推定の平均適合率を示す．提案手法が最も高い精度を示した論文に関しては，数値を太字で示す．各手法の全体での平均適合率は，提案手法を用いた場合に 86 %，比較手法 1 を用いた場合約 79 %，比較手法 2 を用いた場合約 79 %という結果が得られ，提案手法が最も高い平均適合率を示した．

表 2 に，各手法の平均適合率についての優劣関係を示す．比較手法 1 に対しては 54%，比較手法 2 に対しては 42%の論文において，提案手法は比較手法に比べて高い精度を示した．比較手法が提案手法に比べて高い精度を示した論文は，比較手法 1 で 17%，比較手法 2 で 29%であった．提案手法は，概ねの論文において比較手法に比べて高い精度で重要な図表を抽出可能であることが示唆された．

キャプション文がある程度以上の文量があるものについては，比較手法 1 が有効であったと考えられる．しかしながら，今回対象とした論文内では“提案手法”や“実験結果”など極端に短い文でキャプション文が構成されているものも多く見られたため，論文毎に精度のばらつきが見られたと考えられる．これに対し，提案手法で用いた図表引用文と周辺文の位置関係は全ての論文で共通の特性であるため，論文によらず高い精度を示したと考えられる．比較手法 2 は，文と図表との関連を一切考慮していない．論文では，文章を主要な情報として，その追加情報や補助情報として図表を使うことが多く，文と図表には少なからず何かしらの関連性があると考えられる．比較手法 2 に比べて提案手法が高い精度を示したことで，図表の大きさのみならず，図表に関連した文の重要度を考慮することが有効であることが示唆された．これらの結果より，「文重要度を図表に伝搬する際に図表引用文と各文の位置情報を用いること」および「文重要度を図表に伝搬すること」の有効性がそれぞれ確認された．

表 1 各手法を用いた場合の論文ごとの平均適合率

論文 ID	提案手法	比較手法 1	比較手法 2
1	0．982	0．962	1．000
2	1．000	1．000	1．000
3	**1．000**	**0．877**	**0．877**
4	1．000	1．000	1．000
5	0．716	0．909	0．863
6	0．886	0．689	0．957
7	**0．645**	**0．340**	**0．474**
8	**0．700**	**0．639**	**0．478**
9	1．000	1．000	1．000
10	0．700	0．533	0．756
11	1．000	0．958	1．000
12	0．750	0．833	0．333
13	1．000	1．000	0．679
14	0．768	0．668	0．830
15	0．725	0．728	0．760
16	0．287	0．288	0．392
17	**1．000**	**0．417**	**0．500**
18	1．000	1．000	1．000
19	**0．796**	**0．786**	**0．760**
20	1．000	1．000	1．000
21	**0．982**	**0．909**	**0．755**
22	**0．796**	**0．660**	**0．596**
23	**0．909**	**0．826**	**0．874**
24	1．000	1．000	1．000
平均	0．860	0．793	0．787

5.3 考察

具体的な事例に注目して，実験結果を考察する．どちらの比較手法に対しても提案手法が高精度を示した論文例として論文 ID17 の論文 [9]，提案手法が特に精度を欠

[1] http://www．ai-gakkai．or．jp/jsai2016/

表 2　各比較手法との平均適合率の優劣関係による論文の割合

	比較手法 1	比較手法 2
提案手法＞比較手法	54 % (13)	42 % (10)
提案手法＜比較手法	17 % (4)	29 % (7)
提案手法＝比較手法	29 % (7)	29 % (7)

表 3　論文 ID17[9] での重要度推定の結果と図表引用回数. 括弧内には正規化前の図表重要度 $CIMP$ を示す.

図表番号	推定重要度	引用回数
図 2	**49 % (10393)**	**4**
図 1	**27 % (5694)**	**3**
図 3	12 % (2468)	1
図 4	8 % (1608)	1
図 5	4 % (915)	1

表 4　論文 ID16[10] での重要度推定の結果と図表引用回数. 括弧内には正規化前の図表重要度 $CIMP$ を示す.

図表番号	推定重要度	引用回数
図 9	15 % (10022)	4
図 10	12 % (7911)	3
図 5	10 % (6945)	2
図 8	10 % (6856)	2
図 1	8 % (5440)	2
図 11	**8 % (5344)**	**2**
図 12	**8 % (5249)**	**2**
図 7	7 % (5162)	2
図 3	**7 % (5138)**	**2**
図 6	6 % (4224)	1
図 2	**5 % (3112)**	**1**
図 4	4 % (2645)	1

いた論文例として論文 ID16 の論文 [10] を挙げる．表 3 と表 4 にそれぞれ，論文 ID17 の論文と論文 ID16 の論文における提案手法による重要度推定の結果と各図表の引用回数を示す．太字となっている図表番号がポスター中で用いられていた本実験における正解図表を示す．なお，括弧内の数値は正規化前の推定重要度 $CIMP$ を示しており，重要度の降順に並び替えて示している．

提案手法では，論文中で複数回引用されている図表に関して，各引用文との位置情報を用いて重要度を算出し，その総和を図表の重要度として推定を行う．そのため，引用回数が多い図表は比較的重要度が高めに推定される傾向にある．表 3 に示す論文のように論文中での引用回数が多い図表がそのままポスター中で用いられるといった，引用回数の多少が図表重要度に影響している論文に関しては比較的有効に働くと考えられる．一方で，表 4 に示す論文のように論文中での引用回数が多いからといって必ずしも図表の重要度は高くないような，引用回数の大小が図表重要度に影響していない論文に対しては提案手法は有効でないと考えられる．

それぞれの論文中での $CIMP$ に着目してみる．論文 ID17 では重要度が最小と推定された図 5 から最大と推定された図 1 にかけて段階的に $CIMP$ が高くなっていく．一方で，論文 ID16 では，最大の重要度をもつと推定された図 9 は他の図に比べて高い値を持つものの，表 4 における図 10〜図 3 に対してはほぼ横ばいの数値を示している．インタフェースなどの例示として多くの画像を示した論文については各画像に対しての重要度に大きな差異が見られない．一方で，手法の中身を補助的に説明するために画像を用いている論文では，図表の重要度に差異が見られやすいといった傾向が示唆される．これらは論文中での図表掲載の意図の違いにもよるものと考えられる．例えば，本稿では着目しなかった数式に対しての評価値を取り入れることで，今後は図表の重要度のみならず，図表の役割についても考慮していく．

6　おわりに

本稿では，論文中の図表の重要度を推定する手法を提案した．提案手法では，単語の出現頻度と文同士の類似度を元に算出した文重要度を，図表引用文との位置情報を用いて図表へ伝搬することで図表の重要度を推定した．評価実験では，キャプション文を用いて文重要度を伝搬する手法および論文中での図表の大きさを用いる手法に比べて提案手法が高い精度を示した．一方で，論文中での引用回数が少ないにも関わらず重要な図表に関しては，現在の提案手法では重要度の推定が難しいという課題も発見された．重要度推定において図表の引用回数の考慮，数式に基づいた図表の推定される役割の考慮など，今後検討していく．

本研究は，Yuting Qiang ら [11] によって提案されている論文からポスターを作成するシステムと組み合わせることを展望としている．Yuting Qiang らのシステムでは，論文からポスターに使用する図表を人手で選んでおり，本稿の手法と組み合わせることでポスター作成の完全自動化が期待される．また，ポスター中での図表サイズを図表重要度を用いて決定するなどの図表重要度を用いた応用も考えられる．

その他，本稿では論文を画像情報と言語情報が組み合わさって成り立っているものとして扱ったが，コミックや絵本なども論文と同様に画像情報と言語情報が組み合わさって成り立っているものであると考えられる．これ

らの他ドメインについても本研究のアイデアを応用する可能性を探っていく．

謝辞

豊橋技術科学大学の吉田光男助教には，本研究に対して様々な関連研究等アドバイスを賜った．また，本研究では JSAI2016 のポスターセッションにて発表された研究発表論文およびポスターの情報を実験・分析に使わせて頂いた．各著者に記して謝意を示す．

参考文献

[1] 三浦宏一，浜田玲子，井手一郎，坂井修一，田中英彦，“動きに基づく料理映像の自動要約手法,” 画像の認識・理解シンポジウム（MIRU2002）論文集，vol.2，pp.203-208，2002.

[2] T. Yao, T. Mei, and Y. Rui, “Highlight detection with pairwise deep ranking for first-person video summarization,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.982-990, 2016.

[3] 笠松沙紀，伊藤貴之，“動画像データの要約可視化インタフェースの一手法,” DEIM Forum 2009，pp.E9-5，2009.

[4] 市野順子，箕牧数成，山口和泰，垣智，東郁雄，古田重信，“図表検索のための図表情報自動抽出の試み,” 情報処理学会研究報告，no.28，pp.143-150，2002.

[5] 竹島亮，渡邉豊英，“文と単語の相互依存性に注目した図表説明文の抽出,” 電子情報通信学会技術研究報告，vol.110，no.85，pp.43-48，2010.

[6] K. Zechner, “Fast generation of abstracts from general domain text corpora by extracting relevant sentences,” Proceedings of the 16th conference on Computational linguistics, vol.2, pp.986-989, 1996.

[7] 阿辺川武，難波英嗣，高村大也，奥村学，“機械学習による科学技術論文からの書誌情報の自動抽出,” 情報処理学会研究報告，pp.83-90，2003.

[8] “Mecab: Yet another part-of-speech and morphological analyzer,” http://taku910.github.io/mecab/.

[9] 熊谷沙津希，伊藤貴之，本橋洋介，梅津圭介，高塚正浩，“クラスタリングとヒートマップによる高次元データ可視化,” 2016 年度人工知能学会全国大会（第 30 回），pp.1E4-4in2，2016.

[10] 佐野正和，増田英孝，山田剛一，福原知宏，“陸上競技ブログからの活動記録抽出と可視化,” 2016 年度人工知能学会全国大会（第 30 回），pp.1D2-1in2，2016.

[11] Y. Qiang, Y. Fu, Y. Guo, Z.H. Zhou, and L. Sigal, “Learning to generate posters of scientific papers,” Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI’16), pp.51-57, 2016.