SNS 上のタグ付き写真データセットからの語間関係抽出

前西　鷹†　田島 敬史††

†,†† 京都大学大学院情報学研究科　〒 606-8501 京都府京都市左京区吉田本町
E-mail: †maenishi@dl.kuis.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　検索において，検索ワードは結果の絞り込みに利用されていると考えられるが，その能力は語ごとに異なっている．例えば，画像検索エンジンを用いて写真を検索する際，写真の内容をより詳しく限定するために二つの語を用いて AND 検索を行うことがある．このとき，二つの検索ワードが「名詞＋形容詞」や「上位語＋下位語」などの組み合わせの場合，一方の語が写真の内容を大きく決定づけ，もう一方の語はそれをさらに絞り込む働きをしている．このように，語には検索結果の内容を絞り込む能力，すなわち，内容に与える情報量の大きさが，語そのものの性質として備わっているものと考えられる．本研究では，写真の検索において語が持つこのような性質に着目し，AND 検索の結果が各々の語によってどの程度限定されたかを求めることで，語同士の上位・下位関係などの概念を抽出する手法を提案する．データセットには，代表的な写真共有 SNS である Instagram のデータを用いる．最後に，実データに対して提案手法を適用した場合の性能評価と，応用例について述べる．実験の結果，写真内容に情報量を与える語の推定が妥当な精度で行えていることが確認できた．

キーワード　画像検索, SNS, Instagram, 写真, タグ, アノテーション, 語間関係, 関係抽出

1. はじめに

二つの語を検索ワードとして写真を検索を行うとき，検索者はどのような結果を期待しているだろうか．もちろん，二つの語の両方が内容に反映されている写真を求めていることには間違いないが，これらは二つのパターンに分類できることが多い．一つ目は，検索ワードの両方が写っている写真を求めているという場合であり，もう一つは，一方の語をもう一方の語で修飾し，より限定的な写真を求めているという場合である．

前者の場合は「sea sky」やなどが，後者の場合は「woman cute」などが検索ワードの例として考えられる．多くの場合，前者を検索ワードに用いた画像検索の結果は，海と空がそれぞれ両方写った写真の集合となり，後者を検索ワードに用いた画像検索の結果は，可愛らしい女性が写った写真の集合となるだろう．Google Images(注1) を用いて検索を行った結果の一部を図 1 に示す．

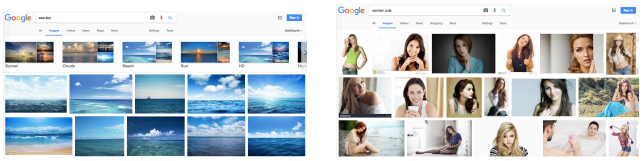

図 1: 「sea sky」の画像検索の結果と,「woman cute」の画像検索の結果

この違いは語そのものの性質に起因していると考えられる．前者の「sea」「sky」はそれ自身が写真に写るものを表す，すなわち，写真に写るものを決定する語であるが，後者の「cute」は「woman」を修飾する語であり，それ自身が写真に写るものを決定することはない．つまり,「sea」「sky」のような語は,「cute」のような語に比べて，写真の内容を決定する（絞り込む）能力，すなわち，写真の内容に与える情報量が大きい語であるといえる．

この性質は，それぞれの語一つだけを検索ワードに用いて画像検索を行うと明らかである.「sky」や「woman」などは内容に与える情報量が大きい語であるため，それ自身を検索ワードに用いて画像検索を行った結果は絞り込まれている．一方で,「cute」などは内容に与える情報量が小さい語であるため，それ自身を検索ワードに用いて画像検索を行った結果はあまり絞り込まれておらず，様々な内容の写真が得られる（図 2）．

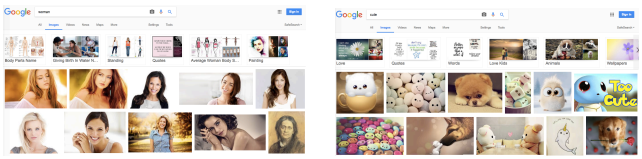

図 2: 「woman」の画像検索の結果と,「cute」の画像検索の結果

このように，語が写真の内容に与える情報量は語によって異なり，そしてそれは語そのものが性質として有している概念であると考えられる．最も単純には，名詞の語はそれ自体が物を表すため，写真の内容を決定づける性質を持っていることが多いと考えられるが，形容詞は他の名詞を修飾する語にすぎず，それ自体が写真の内容を決定づける性質は大きくないことは予想できるだろう．しかしながら，名詞には物を表す語だけではなく，時間などの抽象的な概念を表す語なども多く，同じ名詞でもその性質の大きさは様々であるため，品詞のみからこれを求めることは難しい．それゆえ，次のようなアプローチをとる．

(注1)：https://images.google.com/

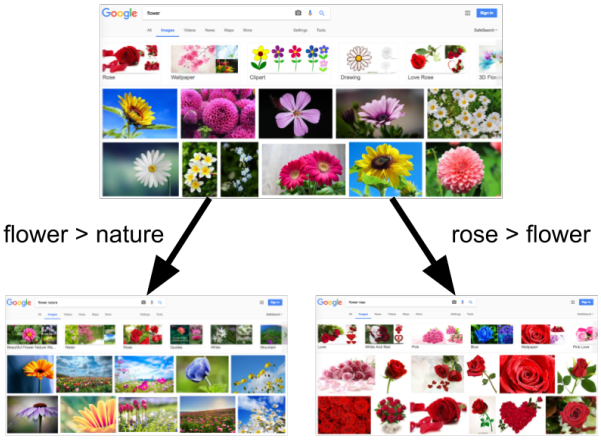

図 3: 「flower」に「nature」と「rose」を加えて AND 検索を行うことによる結果の変化

ここで，一つの語だけを検索ワードに用いた場合と，そこに別の語を加えて AND 検索を行った場合の結果を比較すると，後から加えた語がもとの語よりも多くの情報量を内容に与える語であった場合，写真に写ってるものは大きく変化するが，そうでない場合は写真に写っているものはあまり変化しないことがわかる．

例えば，「flower」に，「nature」を加えて AND 検索を行っても，結果はやはり（いろいろな種類の）花の写真であることには変わりないが，「rose」を加えて AND 検索を行うと，結果は（花の中でも特に）バラの写真となり，写真に写っているものが変化する．これは，語が内容に与える情報量の大きさが nature < flower < rose であることによる（図 3）．

このとき，内容に与える情報量の大きさの大小関係が，それぞれの語の概念上の上位・下位関係を表していることに注目したい．二つの語で AND 検索を行った結果と，二つの語をそれぞれ一つだけで検索した場合の結果を比較することで，二つの語のどちらがより上位の概念を表す語であるかを求めることができる．この性質は品詞に依存するものではないため，それぞれが全て名詞である nature, flower, rose の三語に対しても，正しく上位・下位概念が現れている．

本研究では，ここまでで述べた性質に着目し，語が写真の内容に与える情報量の大小関係という観点から，語の概念上の関係を推定することを考える．これにより，品詞という枠組みでは捉えることのできない語間関係を抽出することが可能となる．

本論文で提案する手法を大量のデータセットに適用することで，同義関係や上位下位関係などの概念を蓄積した辞書作成へも応用が可能と考えられる．さらに，データセットに SNS 上のデータを選ぶことで，常に更新され続ける概念辞書の構築が可能となる．このような応用が期待できる点を，本研究の主な貢献としたい．

2. 関 連 研 究

本章では，既存の関連研究を紹介するとともに，本研究のアプローチにおいて重要となる概念や考え方について述べる．

図 4: Instagram の投稿

2.1 アノテーション

アノテーションとは，データに対して注釈となる情報をメタデータとして付与することであり，また，付与されたデータそのものを指す場合もある．

本研究では，ある語で検索した結果を，その語が付与されたデータの集合であるとみなす．例えば，「flower」で検索した結果得られる写真の集合は「flower」がメタデータとして付与された写真の集合であり，「flower nature」で検索した結果得られる写真の集合は「flower」と「nature」の両方が付与された写真の集合である．ただし，検索ワード以外の語が付与されているものも検索結果には含まれている．

このようにしてメタデータを付与することを，しばしば**「タグ付けする」**とも呼ぶ．最近では，文献 [1] など，画像データに対して自動でアノテーションを行う研究も盛んに行われている．

2.2 Instagram

本研究ではデータセットに Instagram(注2) の投稿データを用いる．Instagram は 2010 年にサービスを開始した画像共有 SNS であり，2016 年 12 月にはユーザー数が全世界で 6 億人を突破(注3) するなど，世界中で最も利用されているサービスの一つである．

スマートフォンで撮影した写真を綺麗に加工して共有できる点などから，若い世代を中心に非常に人気が高く，トレンドが非常に現れやすいという性質がある．それゆえに最近ではマーケティングの分野でも盛んに利用されている．

Instagram の投稿の特徴の一つに，前節で述べたタグ付けがある．ユーザーは写真を投稿する際，投稿内容を説明するためのタグを付与する．Instagram はタグ付けが非常に盛んな SNS であり，一つの投稿に数十個のタグが付与されることも珍しくない．図 4 に，Instagram の投稿の一例を示す．

したがって，Instagram のデータはアノテーションが行われた写真データであるため，Instagram 上でタグを用いて投稿を検索して得られたデータは，画像検索で得られたものと同等に扱うことができる．検索エンジンではなく Instagram からデータを得ることで，トレンドを反映しやすい性質などから，最新

(注2)：https://www.instagram.com/
(注3)：http://blog.instagram.com/post/154506585127/161215-600million

でかつ信頼度の高いデータが得られることが期待できるというメリットがある．

なお，Instagram のタグ付けは投稿者が完全に自由に行うことができるため，必ずしも内容を完璧に説明したタグ付け（**「タグ付けが理想的である」**とよぶ）が行われているわけではない．しかしながら，文献 [2] では，画像に自動でタグデータを付与するための機械学習に Instagram のデータを用いた結果，十分機械学習に利用できると報告している．すなわち，Instagram のタグデータは写真の内容を十分精度良く表現していることが示されている．

また，文献 [3] では，Instagram と同様に，タグ付けを行うことができる写真共有 SNS である Flickr(注4) において，タグ同士の共起度や，画像の持つ特徴量などをもとに，タグ付けされたそれぞれの語が写真内容とどの程度一致しているかを推定し，画像検索に応用している．本研究での提案手法も画像検索に応用可能であると考えており，この点で共通している．

本研究で利用する Instagram 上のデータには，付与されたタグデータ以外にも写真そのものも含まれているが，本研究では色などの画像の特徴量は利用しない．また，語のタグ付けは理想的である，すなわち写真の内容を表すために必要な語はすべてタグ付けされていると仮定して問題を設定している．

2.3 自然言語処理

前節でも述べた通り，本研究では画像そのもののの特徴量は一切利用しておらず，利用しているのはタグ付けされた語のデータのみであるため，実際に扱うのは単語である．文書を単語に分割するのと全く同様のアプローチを行っているため，自然言語処理の分野でよく使われる手法も利用している．本節では，中でも本研究と関連の深い手法について述べる．

2.3.1 ベクトル空間モデル

ベクトル空間モデルとは，主に文書検索の分野で利用される，情報検索や情報推薦を行うためのアルゴリズムの一種であり，文書同士の類似度計算などを目的に，文書や質問文中に登場する単語，その出現回数を高次元のベクトルの形で表現する．ベクトル空間モデルについては詳しくは文献 [5] などを参照されたい．

2.1 節で述べたように，検索ワードを用いて検索した結果は，その語がタグ付けされているデータの集合であると考えられる．本研究では，写真の集合を，それらの写真にタグ付けされた語の集合ととらえベクトル空間モデルを用いて表現し，写真集合の類似度計算や，タグ付けされた語同士の共起度計算を行う．詳しくは，3. 章で述べる．共起語集合の類似度を利用した既存研究には文献 [4] などがある．

2.3.2 分布類似度

分布類似度とは，似たような意味の語は似たような文脈で利用されるという仮定に基づき計算される語の類似度である．本研究でもこれと同様の仮定に基づき，似たような意味の語には，似たような語が同時にタグ付けされる（共起する）ことが多いと考える．

分布類似度については，文献 [6] などを参照されたい．

(注4)：https://www.flickr.com/

3. 提案手法

本章では，1. 章で述べたような，語が写真内容に与える情報量の大小関係を写真データから抽出するための具体的な手法について述べる．

3.1 データセット

2.2 で述べた通り，本研究では Instagram の投稿データを利用する．投稿データにおいては，タグ付けされた語のみをデータとして扱い，写真の色情報などの画像特徴量は利用しない．これは，タグ付けが理想的であり，写真の内容がタグ付けされた語で完全に表現されているという仮定に基づいている．

また，必要なデータの収集においては，Python で書かれたクローラを利用し，投稿の言語は英語に限定した．

3.2 アルゴリズムの概要

語 A に対して，A が写真内容に与える情報量の大きさを $I(A)$ で表す．また，語 A,B から作成された二つのベクトル $\vec{A}, \vec{B}$ に対して，$\vec{A}$ と $\vec{B}$ の類似度を $sim(\vec{A}; \vec{B})$ で表し，これを計算するための関数を**類似度関数**とよぶことにする．

このとき，語 A,B に対して，$I(A)$ と $I(B)$ の大小関係を求める流れは次のようになる．

（1） 語 A,B でそれぞれ検索した結果得られる写真の集合を，写真にタグ付けされた語の集合であるとみなしてベクトル空間モデルで表現し，$\vec{A}, \vec{B}$ を構成する．

（2） 同様にして $\overrightarrow{A\&B}$ を構成する．

（3） $sim(\vec{A}; \overrightarrow{A\&B})$ と $sim(\vec{B}; \overrightarrow{A\&B})$ をそれぞれ計算し，次に示す関係から，$I(A)$ と $I(B)$ の大小を導く．

$I(A) > I(B)$
$\Leftrightarrow$ 語 B よりも語 A のほうが，写真の内容に与える情報量が大きい，すなわち，写真の内容を決定付ける度合いが大きい．
$\Leftrightarrow$ A&B による検索の結果得られる写真は，語 B よりも語 A によって内容が説明されている．
$\Leftrightarrow$ $\vec{B}$ よりも $\vec{A}$ のほうが $\overrightarrow{A\&B}$ に類似している．
$\Leftrightarrow$ $sim(\vec{A}; \overrightarrow{A\&B}) > sim(\vec{B}; \overrightarrow{A\&B})$

ここまでの議論からこの関係が成立することが期待される．この手法を用いることで，(AND 検索の結果が十分に存在する) 任意の二つの語 A,B に対して，$I(A), I(B)$ の大小関係を求めることが可能である．

3.3 語ベクトルの具体的な構成方法

本節では，写真集合をベクトル空間モデルを用いて表現する方法について詳しく説明する．

例えば，語「flower」で検索した結果，(1)〜(3) に示す三つ

(1)

(2)

(3)

(1)	flower, nature, beautiful
(2)	rose, flower
(3)	sunflower, beautiful, flower, sky

表 1: それぞれの写真にタグ付けされていた語

	flower	nature	beautiful	rose	sunflower	sky
(1)	1	1	1	0	0	0
(2)	1	0	0	1	0	0
(3)	1	0	1	0	1	1

表 2: それぞれの写真をベクトル空間モデルで表現した結果

	nature	beautiful	rose	sunflower	sky
$\overrightarrow{flower}$	1	2	1	1	1

表 3: flower ベクトルの構成結果

の写真が得られたとする．また，それぞれの写真には，表 1 で示されるようなタグが付けられていたとする．

ベクトル空間モデルを用いると，まずそれぞれの写真は，表 2 のように表現される．

そして，集合内に含まれる全ての写真を，それにタグ付けされた語を用いたベクトル空間モデルで表現し，それら全ての和をとったベクトルを，この集合全体を表現する**語ベクトル**と定義する．ただし，全ての写真に flower がタグ付けされていることは自明であるため，集合全体を表す語ベクトルにはそれ自身は含まないものとする．

これらを踏まえると，flower ベクトルは結局表 3 のように表現される．

もちろん,「flower」というタグがついているからといって，必ずしもこのような写真ばかりであるとは限らない．例えば，以下に示す (4) のような投稿は花柄のドレスの写真であり，写真の内容に最も情報量を与えている語は「flower」ではなく「dress」であり,「flower」はここでは「dress」の内容を限定する働きをしている．

(4)

タグ：
dress, flower, fashion, beautiful

このような写真は「flower」という語だけで検索した場合の正解ではない．このことは，次のように定義できる．

語	値（/1000）
flowers	248
nature	242
love	139
beautiful	131
flowerstagram	106
garden	95
summer	89
rose	86
photography	70
spring	64

表 4: flower ベクトル（上位 10 件を抜粋）

写真 p が，語 A での検索結果における正例である
$\Leftrightarrow$ p の内容に最も情報量を与える語が A である．なお，最も情報量を表す語が二つ以上ある場合も正例に含むものとする．

表 3 では説明のためデータ数は 3 としたが，データ数をさらに増やすことで，語ベクトルはその語と同時にタグ付けされる割合（**共起度**）が大きい語の統計的な集計となる．これにより，語ベクトルはその語と共起度の高い語（**周辺語**とよぶ）によって特徴付けられる．なお，共起度は厳密には以下のように定義する．

共起度
語 A がタグ付けされた投稿 n 件中に，語 B が同時にタグ付けされた投稿が k 件あるとき，語 A に対する語 B の共起度を k/n で定義する．

データ数を 1000 とした場合の flower ベクトルを，成分の大きい 10 語について抜粋して表 4 に示す．

このベクトルが表している事柄としては，flower がタグ付けされた写真 1000 件のうち，flowers も同時にタグ付けされていたものは 248 件，nature も同時にタグ付けされていたものは 242 件，…ということである．このように，語ベクトルの主成分は，その語と共起度の大きい周辺語で構成され，これによってもとの語が特徴付けられていることがわかる．

なお，表 4 中にみられる「flowerstagram」という語は flower+instagram から生まれた造語である．タグ付けの自由度が高いがゆえにこのような造語が非常に多いことも，Instagram の特徴の一つである．

ここまでは，flower など写真の内容に与える情報量が大きい語について考えていたが，形容詞や抽象名詞など，そもそも内容に与える情報量が小さい語でベクトルを構成する際は解釈が少し異なる．例として，データ数を 1000 とした場合の nature ベクトルを，先ほどと同様に成分の大きい 10 語について抜粋して表 5 に示す．

ベクトルの構成方法から，nature と共起度の高い周辺語で構成されているという点は同じであるが，nature という語が表す

語	値（/1000）
photography	140
beautiful	139
travel	136
sky	123
love	113
instagood	113
landscape	104
photooftheday	99
summer	98
sunset	96

表 5: nature ベクトル（上位 10 件を抜粋）

語	相互共起度
flowerstagram	0.067
flowers	0.062
bloom	0.047
floral	0.029
blossom	0.028
flor	0.025
rose	0.024
petals	0.023
nature	0.022
flowersofinstagram	0.018

表 6: 相互共起度を成分とした flower ベクトル（上位 10 件を抜粋）

概念が flower に比べると少し抽象的であるため，写真の内容に与える情報量は少し小さい．このような語については，そのベクトルに現れる語もまた，写真の内容に与える情報量が小さいものが多くなる．

また，nature などの語で写真を検索した場合に得られる写真は，すべて nature の下位概念を表すものが写った写真である．定義に従うとこれらはすべて負例となるが，このような語で検索する場合には，そもそも検索者が多様な結果を求めていると考えられるため，正例とも負例ともいえない．

3.4 相互共起度に基づくベクトルの更新

語 A のベクトルにおいて，ある語 B の成分が大きくなる（語 A に対する語 B の共起度が高くなる）要因としては，以下の二つが主に考えられる．

（1） 語 A と語 B が概念的に近く，よく同時にタグ付けされるため．

（2） 語 B がよく使われる語かつ写真の内容に与える情報量が少ない語であり，さまざまな内容の写真によくタグ付けされるため．

例えば，表 4 の flower ベクトルにおいて,「flowers」や「nature」などといった語は,「flower」と概念的に非常に近く，これらもまた「flower」で表される写真の内容を表す語であるため，(1) の要因に合致する．

一方で,「beautiful」や「love」などといった語は，それ自身が非常によく使われる語であり，さまざまな投稿にタグ付けされているため，flower ベクトルに限らずさまざまなベクトル中に現れるが，写真の内容に与える情報量は極めて少ない．これらは (2) の要因に合致し，このような語はベクトルを特徴付ける能力が低いと言い換えることもできる．

語ベクトルは周辺語によって語を特徴付けることを目的として構成されているため，このような語が，ただ共起度が高いというだけでベクトル内で成分が大きくなってしまっているのは相応しくない．したがって，このような語の成分値を下げるために相互共起度という概念を導入する．

相互共起度

語 A と語 B の相互共起度を（語 A に対する語 B の共起度）×（語 B に対する語 A の共起度）で定義する．

これまで用いていた共起度は相互ではなくどちらか一方のみを考えていた．これにより，どんな語ともある程度共起度が高くなるような語がベクトルに含まれてしまう．そこで,「love」のような (2) の要因に合致する語の場合,「flower」に対する「love」の共起度は大きいが,「love」に対する「flower」の共起度は小さいことを利用し，ベクトルの成分をこれらの積で更新する．これにより，ベクトルの内容をより特徴付ける表現が可能になる．このようにして成分を相互共起度で更新した flower ベクトルを表 6 に示す．

「beautiful」や「love」のような語の順位が下がり，flower と概念的に近い語のみが上位に現れているため，一目見て性能が改善されたといえる．以下では，語ベクトルの成分はすべて相互共起度であるとする．

3.5 AND ベクトルの構成

次に，二つの語の AND ベクトルの構成について説明する．基本的な考え方は同じで，AND 検索の結果得られた写真にタグ付けされた語の集合を，ベクトル空間モデルで表現する．ただし，もとの語が二つ存在するため，それらはどちらも成分に含まない．

表 7 に，flower ベクトル，nature ベクトル，flower&nature ベクトルのそれぞれを，成分の大きい 10 語について抜粋して示す．

この結果を確認すると，flower&nature ベクトルの成分には,「flower」に概念的に近い語が多く含まれており，これは nature ベクトルよりも flower ベクトルに近いことが推測できる．そこで，ベクトルの類似度を実際に計算することで，これを確認する．

3.6 類似度計算

ベクトルの類似度の比較にはいくつか方法が考えられるが，ここでは最もシンプルな手法であるコサイン類似度を用いる手法と，情報量の観点からアプローチをする手法の二つを紹介する．

3.6.1 コサイン類似度による方法

コサイン類似度とは，情報検索の分野で主に用いられる，二つのベクトルの類似度をそれらのなす角 θ のコサインの値で表す手法である．

flower		nature	
flowerstagram	0.026	naturelovers	0.029
flowers	0.024	sky	0.027
bloom	0.018	landscape	0.025
floral	0.012	naturephotography	0.023
blossom	0.011	clouds	0.018
flor	0.010	green	0.017
rose	0.009	outdoors	0.017
petals	0.009	beautiful	0.016
nature	0.009	forest	0.015
flowersofinstagram	0.008	travel	0.014

flower&nature	
flowerstagram	0.010
bloom	0.009
blossom	0.007
petals	0.006
petal	0.004
flowers	0.004
plants	0.004
floral	0.004
flowersofinstagram	0.004
botanical	0.003

表 7: flower ベクトル，nature ベクトル，flower&nature ベクトル（それぞれ上位 10 件を抜粋）

コサイン類似度

二つのベクトル $\vec{A}, \vec{B}$ に対して，それらのなす角を θ としたときのコサイン類似度 $cos(\vec{A}, \vec{B})$ は

$$cos(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{|\vec{A}||\vec{B}|} = cos\theta$$

で定義される．$-1 \leqq cos(\vec{A}, \vec{B}) \leqq 1$ であり，二つのベクトルの向きが完全に一致したときに 1，全く逆方向を向いているときに-1 となる．

ただし，ここではベクトルの成分はすべて正であるため，コサイン類似度が負になることはなく，$0 \leqq cos(\vec{A}, \vec{B}) \leqq 1$ の値をとる．

表 7 のベクトルに対して，類似度関数を $sim(\vec{A}; \vec{B}) = cos(\vec{A}, \vec{B})$ としたときの類似度 $sim(\overrightarrow{flower}; \overrightarrow{flower\&nature})$ と $sim(\overrightarrow{nature}; \overrightarrow{flower\&nature})$ をそれぞれ計算すると，結果は表 8 のようになる．このとき，$sim(\overrightarrow{flower}; \overrightarrow{flower\&nature}) > sim(\overrightarrow{nature}; \overrightarrow{flower\&nature})$ であり，これより $I(flower) > I(nature)$ であると求められる．

$sim(\overrightarrow{flower}; \overrightarrow{flower\&nature})$	0.870
$sim(\overrightarrow{nature}; \overrightarrow{flower\&nature})$	0.217

表 8: コサイン類似度を用いた類似度計算の結果

しかしながら，コサイン類似度は各々のベクトルの成分に突出して大きな共通部分があるような場合に，類似度が大きくなってしまうという性質がある．本研究で扱う問題に対しては，数値の大きな共通要素があるほど類似度が高いとするよりも，ベクトルが全体として似ているほど類似度が高いと評価するほうが適していると考えられる．そこで，そのような性質をある程度みたす第二の方法について考える．

3.6.2 ダイバージェンスによる方法

カルバック・ライブラー情報量（K-L ダイバージェンス）とは，情報理論の分野において，二つの分布の差異を測るために利用される指標である．この指標は，一方が他方からどれほどの情報量を得るかに基づき，得られる情報量が小さいほど差異が小さいと考える．

K-L ダイバージェンス

二つの確率分布 p, q において，分布 p の分布 q に対する K-L ダイバージェンスは

$$KL(p||q) = \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}$$

で定義される．$KL(p||q) \geqq 0$ であり，分布が全く同じだと 0 になる．

まず，これまでに作成したベクトルにダイバージェンスによる方法を適用するために，ベクトルの成分が確率変数となるよう正規化する必要がある．これは単純に，全ての成分をその和で割ればよい．これにより，語ベクトルの各々の成分は，その語が同時にタグ付けされている確率であると解釈できる．

また，定義内に「分布 p の分布 q に対する」とあるように，この指標には向きが存在する．これを類似度として扱うためには，対称な指標へと変換する必要がある．そこで，K-L ダイバージェンスに次のように対称性を持たせた Jensen-Shannon ダイバージェンス（JS ダイバージェンス）とよばれる指標を用いる．

JS ダイバージェンス

二つの確率分布 p, q において，分布 p と分布 q の間の JS ダイバージェンスは

$$JS(p||q) = \frac{1}{2}KL(p||r) + \frac{1}{2}KL(q||r)$$

で定義される．ただし，

$$r = \frac{1}{2}(p + q)$$

であり，$KL(p||q)$ は分布 p の分布 q に対する K-L ダイバージェンスである．

式の形より，JS ダイバージェンスが p, q に対して対称であることは明らかである．さらに，p, q は確率変数であるから，$p_k, q_k \geqq 0$ であり，$\sum_{k=1}^{K} p_k = \sum_{k=1}^{K} q_k = 1$ が成立している．このとき，$p_k \leqq p_k + q_k$ より $\frac{p_k}{p_k + q_k} \leqq 1$ であることに注意すると，

$$
\begin{aligned}
KL(p||r) &= \sum_{k=1}^{K} p_k \log \frac{p_k}{r_k} \\
&= \sum_{k=1}^{K} p_k \log \frac{p_k}{\frac{1}{2}(p_k + q_k)} \\
&= \sum_{k=1}^{K} p_k \log \frac{2p_k}{p_k + q_k} \leqq \sum_{k=1}^{K} p_k \log 2 = \sum_{k=1}^{K} p_k = 1
\end{aligned}
$$

となる．全く同様に $KL(q||r) \leqq 1$ であるため，

$$0 \leqq JS(p||q) \leqq 1$$

がいえる．

したがって，類似度関数 $sim(\vec{A}; \vec{B})$ を

$$sim(\vec{A}; \vec{B}) = 1 - JS(\vec{A}||\vec{B}) \tag{1}$$

と定めると，これは類似度の性質をみたす．

最後に，表 7 のデータに対し，JS ダイバージェンスを用いた類似度計算を適用する．そのために，まずベクトルを確率変数として扱えるよう正規化を行う．その結果を表 9 に示す．

flower		**nature**	
flowerstagram	0.068	naturelovers	0.028
flowers	0.062	sky	0.026
bloom	0.047	landscape	0.024
floral	0.030	naturephotography	0.022
blossom	0.029	clouds	0.017
flor	0.025	green	0.016
rose	0.024	outdoors	0.016
petals	0.023	beautiful	0.016
nature	0.022	forest	0.014
flowersofinstagram	0.021	travel	0.014

flower&nature	
flowerstagram	0.083
bloom	0.075
blossom	0.061
petals	0.056
petal	0.040
flowers	0.039
plants	0.037
floral	0.036
flowersofinstagram	0.034
botanical	0.032

表 9: flower ベクトル，nature ベクトル，flower&nature ベクトル（それぞれ上位 10 件を抜粋）

これらのベクトルに対して，類似度関数 $sim(\vec{A}; \vec{B})$ を (1) 式で定めたときの類似度 $sim(\overrightarrow{flower}; \overrightarrow{flower\&nature})$ と $sim(\overrightarrow{nature}; \overrightarrow{flower\&nature})$ をそれぞれ計算すると，結果は表 10 のようになる．このとき，$sim(\overrightarrow{flower}; \overrightarrow{flower\&nature}) > sim(\overrightarrow{nature}; \overrightarrow{flower\&nature})$ であり，これより $I(flower) > I(nature)$ であると求められる．

$sim(\overrightarrow{flower}; \overrightarrow{flower\&nature})$	0.876
$sim(\overrightarrow{nature}; \overrightarrow{flower\&nature})$	0.550

表 10: JS ダイバージェンスを用いた類似度計算の結果

4. 評価実験

本章では，ここまでで議論した提案手法を実データに対して適用し，その性能を評価する．

4.1 実験方法

複数のタグが付いた Instagram 上の写真データに対して，タグ付けされている語のすべての二つの組み合わせに提案手法を適用し，写真内容に与える情報量の（相対的な）大小関係を推定する．その後，得られた関係を重み付き有向グラフで表現し，重みを掛け合わせながらエッジの向きにスコアを伝搬するアルゴリズムを適用し，大小関係を数値化することで，相対的な大小関係を絶対的な大小関係へと変換する．

グラフは語をノード，大小関係をエッジの重みで表現する．例えば，「nature」と「flower」がタグ付けされた投稿の場合，3. 章での議論より，写真内容に与える情報量の大小関係は $sim(\overrightarrow{flower}; \overrightarrow{flower\&nature})$ と $sim(\overrightarrow{nature}; \overrightarrow{flower\&nature})$ の大小関係で表されていたため，エッジ A → B の重み $W(A \to B)$ を，

$$W(A \to B) = sim(\vec{A}; \overrightarrow{A\&B}), W(B \to A) = sim(\vec{B}; \overrightarrow{A\&B})$$

で定める．ただし，集合 A&B の件数が 100 件に満たないものについては重み 0 とする．なお，類似度計算には JS ダイバージェンスを用いる．表 10 の結果より，この場合のグラフは図 5 のようになる．

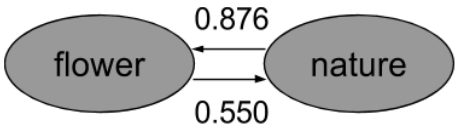

図 5: 「flower」と「nature」がタグ付けされた投稿におけるグラフの構成例

このようにして構成したグラフにおいて，繰り返し回数 k 回のときのノード p_i のスコアの値 $Score(p_i, k)$ を

$$Score(p_i, k) = Score(p_i, k-1) + \sum_{p_j \in M(p_i)} W(p_j \to p_i) Score(p_j, k-1)$$

で定める．なお，$M(p_i)$ はノード p_i に対してエッジが張られているノードの集合であり，各ノードの初期値は $Score(p_i, 0) = 1.0$ と定める．繰り返し回数は $k = 10$ とした．繰り返しの終了後，確率変数として扱えるよう全ノードの値をその和で割って正規化を行ったものを最終的なスコアとする．

今回は，ある語で検索した結果得られた投稿データを 100 件用意し，検索クエリに用いた語がどの程度写真内容に情報量を与えているかを，それぞれの投稿データについて推定し，その

精度を評価する．評価には人手で作成した正解データを用いる．

クエリに用いる語には，Instagram 上でのタグ付け件数が TOP100 の語 (注5) のうち，それ自体が写真の内容を大きく決定づける語であると思われる 5 語「flowers」「dog」「cat」「hair」「girl」を選択し，それぞれについて 100 件の投稿を集め，データセットは合計 500 件とした．

なお，人手での正解データの作成においては，写真とタグ付けされた語の一覧をセットとして与え，それぞれの語が写真内容をどの程度表しているかを 0〜3 の 4 段階で評価した．このような手順で複数人で作成したデータの平均値を正解データとし，正解データにおけるクエリの語のスコアと，提案手法によるクエリの語のスコアとを比較することで精度の評価を行った．具体的には，正解データにおけるクエリの語のスコアを x として，$x = 0, 0 < x \leqq 1, 1 < x \leqq 2, 2 < x < 3, x = 3$ のそれぞれの場合について，クエリの語の提案手法によるスコアの平均値を求め，正解データと推定データの相関度で評価した．

4.2 実験結果

ベースラインには，ナイーブな手法としてタグ付けの順序が先であるほどスコアが高くなるよう重み付けしたものと，比較手法として $W(A \to B)$ を A に対する B の共起度として定めたグラフに同様の PageRank アルゴリズムを適用したものを用いた．

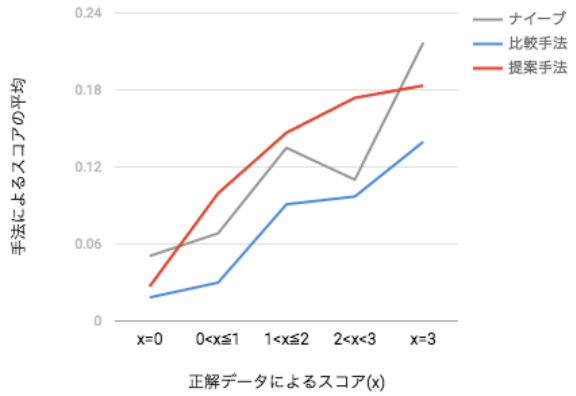

図 6: 実験結果

実験の結果，ベースラインと比較して提案手法による推定値は，クエリの語の正解データによるスコアが高くなるにつれて同様に高いスコアとなっており，最も高い相関が確認された．これは，提案手法による相対的な語間関係の推定と，それを用いて求めた語のスコアがある程度妥当であることを示している．

5. おわりに

本論文では，語が写真の内容に与える情報量は語ごとに性質として備わっており，他の語との相対的な関係で写真の内容が決定づけられることに着目し，その情報量の相対的な大小関係を推定する手法を提案した．

本研究の応用例として，評価実験で述べたような写真内容に対して語がどの程度情報量を与えているかの推定がまず挙げられる．これにより，タグ付けされた語に関するデータから，各々の語が写真の内容をどの程度表しているかが推定でき，写真の内容そのものが推定できる．さらに，これを語と写真内容との一致度とみなすことで，ある語で画像検索を行った際，この一致度が高い順に検索結果を並び替えることで，画像検索のリランキングへと応用が可能である．これは語がタグ付けされた写真データに限らず，アノテーションがなされたオブジェクトに対しても一般的に拡張可能であると考えている．

第二の応用として，二語間の相対的な関係から語間の概念を抽出することが挙げられる．類似度計算に利用するデータを Instagram などの SNS 上のデータとし，最新のデータに対して本手法を適用し続けることで，新たに登場した語や，時代に応じてニュアンスが変化する語に対しても常に最新の情報を得ることができる．これを利用することで，常に新しい概念を含む概念辞書の構築が可能であると考えられる．

これらの応用については今後の課題とするが，このような実用的な応用が期待できる点を，本研究の主要な貢献としたい．

謝辞　本研究は，JST, CREST の支援を受けたものである．

(注5)：http://www.yuiki1994.com/entry/instagram-tag より．データは 2016 年 6 月時点のもの．

文　献

[1] Aaron Duane, Jiang Zhou, Suzanne Little, Cathal Gurrin, and Alan F Smeaton. An annotation system for egocentric image media. In *International Conference on Multimedia Modeling*, pp. 442–445. Springer, 2017.

[2] Stamatios Giannoulakis and Nicolas Tsapatsoulis. Instagram hashtags as image annotation metadata. pp. 206–220, 2015.

[3] Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. Tag ranking. pp. 351–360, 2009.

[4] 梶博行, 相薗敏子ほか. 共起語集合の類似度に基づく対訳コーパスからの対訳語抽出. 情報処理学会論文誌, Vol. 42, No. 9, pp. 2248–2258, 2001.

[5] 後藤正幸, 石田崇, 鈴木誠, 平澤茂一. 高次元ベクトル空間モデルによるテキスト分類問題について：分類性能と距離構造の漸近解析 (理論・技術). 日本経営工学会論文誌, Vol. 61, No. 3, pp. 97–106, aug 2010.

[6] 柴田知秀, 黒橋禎夫. 超大規模ウェブコーパスを用いた分布類似度計算. 言語処理学会年次大会, D4-7, pp. 705–708, 2009.