写真に付与されたタグ群から主題を表すタグの識別手法

前西　鷹†　田島　敬史††

† 京都大学工学部情報学科　〒 606-8501 京都府京都市左京区吉田本町
†† 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　Instagram などの写真投稿型 SNS では，投稿した写真にタグを付与することができ，ユーザーはそのタグを用いて投稿を検索することができる．しかし，写真投稿型 SNS では，一つの写真に多数のタグが付与されることが多く，その中には写真の主題を表す語以外に，主題に関連するその他の語や，写真の印象を表す語なども含まれる．そのため，ある語で検索を行った際，その語が必ずしも主題を表していないような写真も検索結果に含まれることがある．そこで本稿では，タグによる写真検索の適合率向上を目的に，複数のタグが付けられた写真に対して，語の間の意味的関係の情報や，タグ付けの順序情報を用いて，写真の主題を表すタグを識別する手法を提案する．

キーワード　Instagram, 写真, タグ, アノテーション, 主題, 関連語,PageRank

1. はじめに

Twitter や Facebook を始めとする SNS の普及により，ユーザはインターネットを通じて多様な情報をリアルタイムに検索し，入手することが可能になった．SNS から発信される情報はリアルタイム性が極めて高く，企業のアカウントから発信される情報はもちろんのこと，一般ユーザーから発信される情報も有益であることが多いため，SNS に関連した情報検索については，多くの研究がなされている．

Twitter に代表される多くの SNS では，投稿は文章の形で表現されるのが一般的であるが，最近では，Instagram(注1) など写真に特化した SNS も新たに登場している．Instagram は，スマートフォンで撮影した写真を加工してリアルタイムに投稿・共有することができるサービスであり，現在ではユーザー数が Twitter を上回る 4 億人を突破し (注2)，一日あたりの投稿数は平均 8000 万件に達するなど，非常に大きな注目を集めている．そこで本研究では，このような写真投稿型の SNS における投稿の検索に焦点を当てた．

Instagram では投稿に対して，記号#で表されるハッシュタグを付けることができる．例えば，清水寺の写真であれば,「#清水寺」というハッシュタグを付ける．付けられたハッシュタグは投稿の検索に利用され，ハッシュタグを検索条件に指定することで，そのハッシュタグのついた投稿だけを集めて表示することができる．これにより，ユーザーは他のユーザーの投稿を検索することができ，また逆にハッシュタグを自分の投稿に付けることで，より多くの人に自分の投稿を見てもらうことができる．

ハッシュタグ自体は Twitter や Facebook にも存在する機能であるが，Instagram においてはとくにハッシュタグの利用が盛んであり，一つの投稿に対して，10 個以上のハッシュタグがつけられることも珍しくない．Instagram では，ハッシュタグは一つの投稿に上限 30 個までつけることができ，写真に関連のある語を羅列するような形で表現されることが多い．具体的には，写真に写っているもの，写真の撮影地などをまずタグ付けし，その後それらの関連語をタグ付けしていくような付け方がなされる．

具体例として，図 1 にパンケーキの写真を投稿している例を挙げる．

図 1　パンケーキに関する投稿

この投稿にタグ付けされた語のうち，写真に最も主要な被写体として写っているもの，すなわち，写真の主題を最も端的に

(注1)：https://www.instagram.com/
(注2)：“Celebrating a Community of 400 Million”, http://blog.instagram.com/post/129662501137/150922-400million

表す語は何かを考えると,「パンケーキ」(あるいは,「pancake」)であろう．それ以外の語は，写真が撮影された店舗名や地名，パンケーキの関連語など，写真の主題を直接的に表す語ではない．なお，写真の主題については 3.1 で，関連語については 3.3 でそれぞれさらに詳しく議論する．

ハッシュタグを用いてこの投稿を検索することを考える．写真の検索においては,「クエリの語が写真に写っているか否か」または「写真に写っているもの（写真の主題）との関連度が高いか」がクエリとの適合・不適合を判定する基準となることが多い．

例えばこの投稿には「#兵庫」というハッシュタグが付いているため,「#兵庫」で検索した場合にこの投稿は結果に表示されるが,「兵庫」はこの写真が撮影された場所を表す語である「尼崎」の関連語であり，写真に実際に写っている「パンケーキ」とは直接関係がない.「#兵庫」で写真を検索する場合に検索者が結果として期待する写真は兵庫を代表するスポットの写真（例えば，神戸や姫路城など）であることが一般的であり，このような写真は「#兵庫」の検索結果としては適切であるとは言えない．すなわち，投稿には「#兵庫」がタグ付けされているにも関わらず,「#兵庫」で検索した場合の検索結果としては適当でないという事態が起こる．

Twitter の場合，一つの投稿に付けられるハッシュタグは一つ，多くとも二，三であることが一般的であるため，ハッシュタグがついているにも関わらず，ツイート本文とタグ付けの語との関連度が低いという場合は少ない．そもそもハッシュタグは投稿を話題ごとに分類するために考案された機能であり，投稿と全く無関係な語がタグ付けされることは極めて少ないはずである．

しかし Instagram においては，一つの投稿に多くのタグが付けられることと，投稿が文章ではなく写真であることから，写真の主題との関連度が低い語，すなわち，写真の撮影地の情報や，写真に対する印象を表す語などは，投稿者にとっては写真と関係のある語と考えられる語であっても，写真の検索においては有益でない語となってしまうことがある．一つの投稿に対して，写真の主題を表す語とそうでない語が同等にタグ付けされており，ハッシュタグを利用した写真検索の際にはそれらが区別されていないことが問題であると言える．

このように，写真の投稿の検索は，文章で表現された投稿の検索の場合とは異なり，ただ単にタグがついているかどうかだけを調べるだけでは，写真の内容と一致しない結果も表示されてしまうことがあり，適合率の向上のためには更なる考察が必要であると考えられる．

このような問題を背景に，本研究では，ハッシュタグを用いた検索の適合率向上を目的に，写真に複数の語がタグ付けされた投稿において，それらの中から，どの語が写真の主題を端的に表すのかを識別する手法を提案する．提案手法では,「写真の主題となる語には，その関連語が数多くタグ付けされることが多い」という仮定を置き，ハッシュタグのすべての 2 語の組み合わせに対してそれらの関係を抽出する．次に，語の間の関係の中でも特に包含・被包含の関係に着目し，第二の仮定として,「実際に写真に写っているのは包含されている側の語である」という仮定を置く．語間の包含関係は，それらの共起度を用いて推定する．これらの二つの仮定に基づくと，包含される側の語として，最も多くの語と関連を持つ語を主題と推定すればよい．その際，複数の主題を持つ写真もあることや，主題語の同義語が同時にタグ付けされる場合もあることを考えると，ただ一つに主題を決定するだけでは不十分であるため，主題となる可能性の度合いのランキングを求めることを考える．

本論文の構成としては，まず，2. で関連研究について述べる．その後，3. で提案手法を詳しく説明し，関連度の抽出方法や，グラフの具体的な構成方法，グラフを用いたスコアの計算方法について議論する．4. では提案手法を用いた実験とその結果，それらから得られる考察について述べる．

2. 関 連 研 究

本章では，関連研究をいくつか紹介する．

まず，ハッシュタグに関する研究について述べる．ハッシュタグは Instagram だけでなく Twitter でも利用されており，Twitter におけるハッシュタグに関する研究は数多くなされている．Wang らの研究 [4] では，ハッシュタグ同士の関連に基づきグラフを構成し，投稿のクラスタリングを行っており，本研究の手法と非常に似通っている．

次に，Instagram のハッシュタグに関連した研究として，Roi Blanco と Christina Lioma の研究 [2] を挙げる．この研究では，画像に自動でタグデータを付与するための機械学習を行う際，Instagram の投稿が優れたサンプルデータになる，すなわち，Instagram の投稿におけるタグ付けは写真の内容を非常に良く表すデータである，ということを証明している．

Instagram と同様の写真投稿型 SNS のタグに関する研究として，Liu らの研究 [3] がある．この研究では，Flickr について，本研究と同様にタグ付けされた語をノードとするグラフを構成し，ランダムウォークの確率をタグ付け順と語間の共起度を用いて重み付けし，ランダムウォークを反復することで語のスコアを計算している．包含関係を考慮せず，エッジを片方向としている点も本研究とは異なる．

最後に，本研究でのアプローチである「重要な語を識別する」手法の関連研究について述べる．関連研究として，文書から重要語を抽出する研究がある．本研究は文書ではなく語の並びから重要な語を識別する手法であるが，文書から重要な語を識別する手法も，文書から一度重要な語の候補を抽出し語の並びを作成して，その中から重要な語を求めることが殆どである．写真ではなく文書における重要な語という点でも本研究とは少し異なるが，手法においては関連の深い部分が非常に多い．

まず，大澤らの研究 [5] では，文書は筆者独自の考えを主張するために書かれるため，前提となる語が存在するという仮定に基づき，共起度の高い 2 語同士を関連語として抽出し，共起度から Keygraph と呼ばれる単語同士の関連グラフを作成し，重要な語を抽出している．この手法は本研究と手法が非常に似通っている．次に，野本の研究 [7] では,「は」「が」「を」「に」

などの格助詞に着目し主題の候補となる名詞を抽出し，文書との関連度を確率的に求めることで，主題となる語を識別している．同じく野本らの研究 [8] では文書構造に着目し，本文をいくつかのブロックに分け，主題となる見出し語とそれらのブロックの関連度をそれぞれ求めることで，文章中で重要な部分を抽出し，そこから主題を識別している．最後に，Roi Blanco らの研究 [1] と小倉らの研究 [6] では，共起度に基づいて語の関連グラフを構成している点は [5] と同様だが，語に対するスコアの重み付けの計算に PageRank アルゴリズムを用いているという点で，本研究の手法と非常に関連が深い．

3. 提案手法

本章では，1. で述べた提案手法について詳しく議論する．3.1 節で写真の主題についてまず詳しく説明し，3.2 節では 1 語だけに着目した場合の語の性質，3.3 節では 2 語に着目した場合の関係についてそれぞれ議論し，3.4 節で関連語の抽出手法について述べる．3.5 節では，抽出した関係をグラフで表現し，グラフを用いて語にスコアを与える手法について具体的に議論する．

3.1 写真の主題について

まず第一に,「写真の主題」とは何かを考える．写真の主題とは，簡単には「写真に写っている物」であるが，ここでは，次のように定義することにする．

> 語 A が写真 B の主題である
> ⇔ 写真 B に語 A が写っており，かつ，写真 B が語 A をクエリとして写真検索した時の正解である

この定義は，クエリが主題として写真に写っているならば，検索結果に表示されるべきであり，また逆に，検索結果に表示されても問題がないならば，その語は写真の主題である，という考えに基づいている．

しかし，検索結果が正解であるかを判定するのは最終的には人間であることから，正解の判断基準には幅があり，主題は必ずしもただ一つに定まるとは限らない．例えば,「#京都 #清水寺」とタグ付けされたような写真を考えると,「#清水寺」で検索した場合にこの投稿が正解となるのは明らかだが，これが「京都」の写真ではないとはいえず,「#京都」で検索した場合であっても，正解としても問題がない（これは,「清水寺」が「京都」の代表例であることによる）．このような場合には，どちらも正解であると判定したうえで，その適合順位まで正しく求めることを目標とする（この場合は，適合度は清水寺＞京都である）．

もちろん，複数の物が同程度に写っていて，その両方が主題という場合もある．例えば，パンと紅茶が一緒に写った朝食の写真では,「#パン」で検索した場合も,「#紅茶」で検索した場合も，正解と判定するのが妥当であろう．

本研究では，このような写真の主題を識別することを目的とする．

3.2 単体の語の性質

主題となる語を識別するために，語の性質について議論する．まずは，1 つの語だけに着目すると，大きく「主題となりえる語」と「主題になりえない語」とに分類できる．

主題となりえる語は例えば「ケーキ」「夕日」など，具体物を表す語が挙げられる．一方で，主題となりえない語は「オシャレ」「morning」など，具体物ではなく抽象的な概念を表す語が挙げられる（もちろん,「オシャレ」などの抽象的な語で写真を検索する場合もあるが，その場合に正解となる写真の主題は「オシャレ」ではなく,「オシャレなもの」である）．

さらに，主題となりうる語の中でも「主題となる度合い」が存在し，主題となりうる語が複数あった場合は，その度合いが最も大きい語が主題となると考えることができる．例えば,「#京都　#カフェ」とタグ付けされている場合は,「京都にあるカフェ」の写真であると考えるのが自然であり，写真の主題は「カフェ」であると考えられる．ところが，これが「#京都 #カフェ #ケーキ」となると,「京都にあるカフェで食べたケーキ」の写真であると考えるのが自然となり，写真の主題はケーキであると考えるべきである．これは,「主題となる度合い」が「京都＜カフェ＜ケーキ」であるためである．

Google 画像検索 (注1) において,「京都 カフェ」というクエリを与えるとカフェの写真が表示され,「京都 カフェ ケーキ」とクエリを与えるとケーキの写真が表示されることからも，この性質が見てとれる（図 2，図 3）．

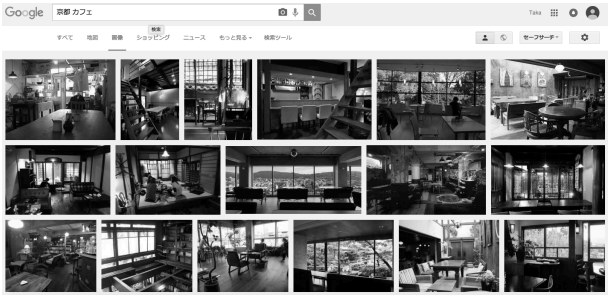

図 2 「京都 カフェ」での Google 画像検索の結果

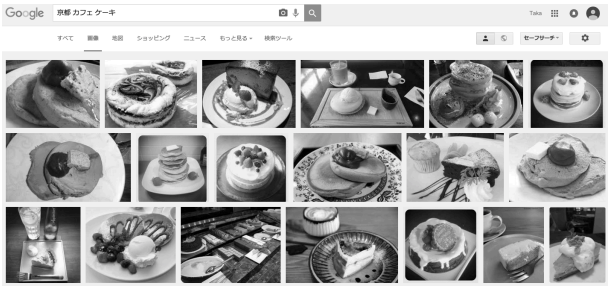

図 3 「京都 カフェ ケーキ」での Google 画像検索の結果

このように，主題となりうる語が真の主題になるかどうかは，単体の語を調べるだけでは判定することはできず，他の語との

(注1)：https://www.google.co.jp/imghp

関係を調べることが必要である．そこで，次節では 2 語間の関係について議論する．

3.3 2 語の関係

次に，2 語の組み合わせに着目する．具体例を挙げながら，2 語がどのような関係にあるのか，その場合に主題となる可能性がより高いのはどちらの語なのかについて，議論する．写真における関係には，以下の (1)〜(4) が考えられる．

(1) 上位・下位関係

「#ケーキ #スイーツ」の場合,「ケーキ」は「スイーツ」の一種であり,「スイーツ」は「ケーキ」の上位概念を表す語である．このような関係にある 2 語を上位・下位関係にあると定義する．このような場合に主題となるのは，概念の下位にある語（この場合は「ケーキ」）である．

(2) 等位関係

「#ケーキ #紅茶」の場合，これらはどちらも写真に写っていて，両方が主題といえることが多い．このように，写真における重要度が同程度である 2 語を，等位関係にあると定義する．「ケーキ」と「パフェ」など，共通の上位語を持つ 2 語は一般に同位関係と定義されるが（この場合は「スイーツ」が共通の上位語），ここでの等位関係はそのような場合に限らず，写真に同程度写る可能性のある 2 語を広く表すものとする．

(3) 同義関係

「#ケーキ #cake」の場合，これらは完全に同じ意味である．このような 2 語を同義関係にあると定義する.「ケーキ」「cake」のような言語の違いによるものだけでなく,「sea」「ocean」のような場合も含む．このような場合は，そのどちらを主題と判定しても問題がない．

(4) 包含関係

「#京都 #清水寺」の場合,「清水寺」は「京都」にあるので,「清水寺」は「京都」に包含されていると考える．このような関係にある 2 語を包含関係にあると定義する．(1) の上位・下位関係が概念的な包含関係を表すのに対し，これは物理的な包含関係を意味し,「#夕焼け #空」のような場合（「夕焼け」は「空」にある）も含むものとする．

これらの結果をまとめると，表 1 のようになる．

関係	A	B	写真の主題
(1) 上位・下位関係	ケーキ	スイーツ	ケーキ
(2) 等位関係	ケーキ	紅茶	両方
(3) 同義関係	ケーキ	cake	どちらとしても問題ない
(4) 包含関係	京都	清水寺	清水寺

表 1　2 語 A,B の関係と，その場合における主題語

(2)，(3) については主題となる可能性が同程度であるが，(1) 上位・下位関係と (4) 包含関係については，関係からどちらが主題となる可能性が高いかを識別できる．これらはそれぞれ概念的・物理的な包含関係を表す関係であり，主題となる可能性が高いのは，いずれも包含関係の内側にある語である．従って，以下ではこれらを区別せず「包含関係」と記すことにし，タグの並びから包含関係を抽出できれば，その最も内側にある語が主題であると決定できるといえる．

そこで次節では，ハッシュタグの並びから包含関係を抽出する方法について述べる．

3.4 包含関係の抽出

語の包含関係の抽出には既存のコーパスを利用する方法も考えられるが，語の辞書的な性質に基づいた包含関係と，写真内に現れる包含関係とは必ずしも一致するとは限らず，また辞書に登録されていない語に対応できないという問題もある．それゆえ，写真における語の関係は，写真に付与されたタグのデータから新たに求めたほうが適当であると考えられる．

そこで，写真に付与されたタグのデータから語の間の包含関係を求めるために，同時にタグ付けされることが多い 2 語を互いに関連語として，関連語を抽出する．これにより，ある語に対して、同時に付けられていることの多いタグを調べれば，その語の関連語を求めることができ，またある 2 語が同時にタグ付けされることが多いかを調べれば，それらが互いに関連語であるかどうかを求めることができる．これらをまとめると以下のようになる．

> （写真において）語 A と語 B に関連がある
> ⇔ A と B が同時にタグ付けされることが多い
> ⇔ A と同時にタグ付けされているタグのランキング上位に B が入っている，あるいは，B と同時にタグ付けされているタグのランキング上位に A が入っている

これによって，既存のコーパスに頼ることなく，与えられたデータに含まれる語の並びから互いに関連のある 2 語を抽出することができる.「上位」を具体的に何位までとするのが最も適切であるかを決定するには更なる実験が必要だが，本研究では 10 位までを上位として考え実験・考察を行った．

次に，関連があると判定された 2 語の組み合わせから，その包含・被包含の関係を求める方法を考える．ここで，具体例として「#京都」と「#清水寺」を考える．この 2 語は同時にタグ付けされることが非常に多く，先に述べた方法で互いに関連があると判定できる．実際に「#京都」がタグ付けされている投稿,「#清水寺」がタグ付けされている投稿をそれぞれ最新 100 件ずつ調べ，同時にタグ付けされている語のランキングを 10 位まで求めると，表 2 のようになった．

「京都」と同時にタグ付けされている語のランキング内に「清水寺」が,「清水寺」と同時にタグ付けされている語のランキング内に「京都」があり，これらは互いに関連語であると判定できる．

ここで，包含関係が「清水寺 ⊂ 京都」であるため，(a)「#京都がタグ付けされた投稿の中で#清水寺がタグ付けされている投稿の割合」と (b)「#清水寺がタグ付けされた投稿の中で#京都がタグ付けされている投稿の割合」を比較すると，(a) < (b) となることに注目する．これは，包含関係から明らかであり，多くの場合逆も成立するため，この関係を用いて包含関係を定

「京都」		「清水寺」	
タグ	件数	タグ	件数
kyoto	46	京都	53
japan	15	kyoto	35
日本	8	japan	23
祇園	6	着物	15
清水寺	6	kimono	12
八坂神社	5	祇園	10
travel	5	地主神社	9
instagood	5	kiyomizudera	9
神戸	4	temple	8
嵐山	4	日本	8

表 2　「#京都」「#清水寺」とそれぞれ同時にタグ付けされている語のランキング

義する．すなわち，関連があると判定された 2 語 A，B に対して，以下が成立する．

（写真において）A ⊂ B である
⇔ A と B が互いに関連語であり，かつ (a)「A がタグ付けされた投稿の中で B がタグ付けされている投稿の割合」と (b)「B がタグ付けされた投稿の中で A がタグ付けされている投稿の割合」を比較した場合に，(a) ＜ (b)

この手法では，同時にタグ付けされることの多い 2 語の関係が包含関係であると仮定している．しかし，同時にタグ付けされることが多いことから抽出できる 2 語の関係には，包含関係だけでなく同義関係も含まれることに注意したい．理想的には，(a) = (b) となるものを互いに同義関係にあると定義すべきだが，この手法だけでは包含関係と同義関係を区別することは困難である．

しかし，3.3 でも述べた通り，2 語が同義関係にあるならば，そのどちらを主題と判定しても問題はないため，同義関係を包含関係と同様に処理しても結果に影響はないと考えた．

なお，等位関係はこの手法では直接抽出できないため，複数の独立した包含関係が出来た場合に，各々の包含関係の内側にある語を等位関係にあると求めることとする．

3.5　関係のグラフ化とスコア計算

3.4 で，包含関係を抽出する手法について述べた．この包含関係の情報を用いて，最も多くの関連語を持ち，かつ包含関係の最も内側にある語を主題であると決定する．

このために，抽出した関係をグラフに表す．それぞれの語をノードで表し，語同士をエッジで結ぶことによって 3.4 で抽出した包含関係を表現し，グラフを用いて各ノード，すなわち各語にスコアを与える．

スコアの計算には PageRank アルゴリズムを用いる．ここで,「多くの関連語を持つ語ほど主題となる可能性が高い」ことを表現するために包含関係 A ⊂ B をエッジ A ← B で表し,「主題となる語の同義語も同様に主題となる」ことを表現するために同義関係 A = B をエッジ A ⇔ B で表す．

1. で挙げたパンケーキの投稿に対して，この手法を用いてグラフを構成すると，図 4 のようになる．

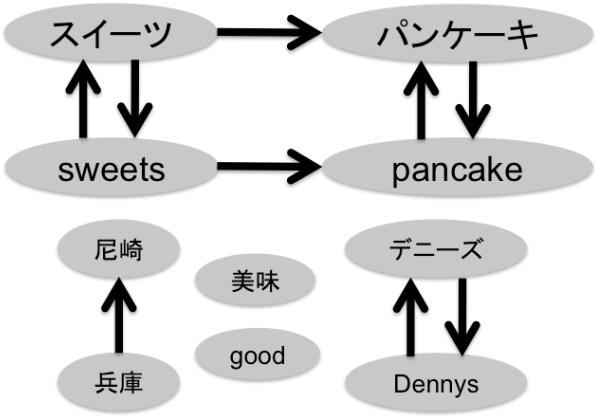

図 4　図 1 の投稿に付けられたタグの語間関係を表現したグラフ

さらに，語同士の関連の度合いを考慮するために，エッジに重み付けを行う．重み付けの基準には，包含関係の判定に利用した「同時にタグ付けされる割合」を用いる．

関連のある 2 語間のエッジの重み付け
エッジ A → B の重み α を「B がタグ付けされた投稿の中で A がタグ付けされている投稿の割合（B に対する A の共起度）」で定義する．(0≦ α ≦1)

例えば，B がタグ付けされた投稿 100 件中，A が同時にタグ付けされている投稿の件数が 30 件であったならば，重み α は 0.30 となる．

さらに,「タグ付けの順番が先であるほど主題となる可能性が高くなる」という仮定に基づき，PageRank におけるランダムジャンプの確率をタグ付けの順序で重み付けした．

タグ付けの順番が前から n 番目の語に対するランダムジャンプの確率 d_n の値を，

$$d_n = \frac{(1-\varepsilon)^n}{\sum_{k=1}^{N}(1-\varepsilon)^k} d$$

で定める．ただし，d：ランダムジャンプ全体の確率，ε：パラメータ，N：全タグの個数である

パラメータ ε を用いて重み付けを行い，正規化を行っている．この重み付けは「主題となる度合いがタグ付けの順番にどの程度依存するか」を表す尺度となり，グラフのエッジの重み付けの方法とともに検討の余地が残されている．今回は $\varepsilon = 0.01$ として計算した．

これらのパラメータを設定したうえで，グラフに PageRank アルゴリズムを適用し，最終的なスコアが高い語ほど主題とな

る可能性が高いと判定する．図 4 の例に対する PageRank アルゴリズムの実行結果は表 3 のようになった．

順位	タグ	スコア
1	パンケーキ	0.149
2	pancake	0.139
3	スイーツ	0.112
4	Dennys	0.108
5	デニーズ	0.107
6	尼崎	0.102
7	sweets	0.101
8	兵庫	0.101
9	美味	0.096
10	good	0.095

表 3　図 4 のグラフに対する提案手法の実行結果

関連語を持つ「パンケーキ」と，その同義語の「pancake」には他よりも高いスコアが付いており，主題となる可能性が高いと判定できた．また，どの語とも関係が抽出できなかった「美味」「good」に対してはタグ付けの順番でランキングされていることがわかる．

4. 評価実験

本章では，本研究でここまでに提案した手法を用いて，実在するデータに対する評価実験を行い，提案手法の実用性を確認する．今回は，以下のような実験を行った．

4.1 実験概要

実験の概要は以下の通りである．

実験概要
(i) Instagram において 3 個以上のタグが付けられた写真の投稿を対象に，データを 500 件収集する．
(ii) (i) のデータに対し，3 人の評価者により正解となるランキングを人手で作成する．
(iii) (i) のデータに対し，提案手法を実行する．
(iv) (iii) で求めたランキングを，(ii) で求めた正解となるランキングと比較する．

正解となるランキングの作成においては，それぞれの投稿に付けられている全てのタグに対し，次に述べる基準に従ってスコア付けを行う．これを 3 人の評価者によって 3 通り作成し，それらの平均値をとったものを実験に用いる．ユーザーによるタグのスコア付けの基準は以下のように設定した．

ユーザーによるタグのスコア付けの基準
スコアは 0,1,2,3 の 4 段階評価とし，点数は以下の評価尺度に従う．
3 点：そのタグで表現されているものが写真に主要な被写体の一つとして写っている．
2 点：そのタグで表現されているものが写真に被写体の一つとして写っているが，明らかに，より主要な被写体が他にある．あるいは，その被写体の表現として，明らかに，より適切な表現のタグが他にある．
1 点：そのタグで表現されているものが，写真中に被写体の一つとして写っているとは言えないが，間接的にその写真によって表現されていることが明らかである．
0 点：写真だけからは，そのタグで表現されているものとの関連が明らかではない．
ただし，同義語には全て同じ点数を付けるものとする．

これにより，提案手法で求めたランキングの性能を評価する．評価においては二つの尺度を導入した．

一つ目はランキングの 1 位のみに着目し，提案手法で求まったランキングの 1 位が，正解ランキングの 1 位と一致していれば 1，一致していなければ 0 としてスコアを与える手法である（Precision@1）．これは，最初の目的である「主題語を決定する」ということがどの程度達成できたかを測る指標になる．正解データにおいて 1 位が複数存在する場合は，それらは全て 1 位すなわち写真の主題であるため，それらの何れかに一致していれば 1 位と一致していると判定した．

二つ目は，スコア付きのランキングの性能を評価する手法の一つである nDCG を用いる手法である．今回は上位 3 位までのランキングがどの程度の精度で求めることができているか（nDCG@3）を評価した．これらにより，精度が 0 から 1 までの数値で表される．

今回は，単純な手法で作成したランキングとの比較を行った．比較対象として，(1) タグによる検索の検索結果数を基準に作成したランキング と，(2) タグ付けの順序情報のみを基準に作成したランキング をベースラインとして用いた．単純な手法によるランキングの作成方法の詳細は以下の通りである．

(1) タグによる検索の検索結果数を基準に作成したランキング
それぞれのタグで写真検索を行った場合のヒット件数をスコアとし，そのスコアが高い順にランキングを作成する．すなわち，よく使われる一般的な語であるほどスコアが高くなる．
(2) タグ付けの順序情報のみを基準に作成したランキング
タグ数を n とし，投稿におけるタグ付けの順番が先のものから順に $n, n-1, n-2, \ldots$ のようにスコアを与え，そのスコアが高い順にランキングを作成する．

4.2 で用いた 500 件のデータに対し，このような手法でランキングを作成し，同様の実験を行いスコアを算出した．ただし，正解ランキングにおいて全てのタグのスコアが 0 となった場合は正解なしとし，手法の適用の対象外とした．今回の実験では，正解ランキングにおいていずれかのタグに 0 以外のスコアが付けられた有効データ数は 500 件のうち 463 件であった．

4.2 実験結果

実験結果は表 4 のようになった．

	ベースライン (1)	ベースライン (2)	提案手法
Precision@1	0.274	0.314	0.263
nDCG@3	0.571	0.634	0.696

表 4 実験結果

Precision@1 についてはベースラインが提案手法を上回ったが nDCG@3 については提案手法がベースラインを上回った．

最後に，まとめと実験の結果から得られた今後の課題について述べる．

5. まとめと今後の課題

本研究では，写真に付与されたタグが多数ある場合に，写真の主題を表す語とそれ以外の語が混在していることを問題として挙げ，それに対する解決策としてタグ付けされた語同士の関係に着目し，語をノードとしたグラフを構成し語にスコアを与え，主題となる語を識別する手法を提案した．

Instagram は日本だけでなく世界中で広く利用されている SNS であることから，本研究で取り扱う投稿も日本語のものだけに留まらず，世界中のあらゆる言語に対しても通用する汎用性の高い手法にしたいと考えている．そのために，まずは英語の投稿に対する実験を今後行う予定である．

今回の評価実験の結果，nDCG の値は単純な手法であるベースラインを上回る結果となり，ランキングの性能が保証された．今後は関連研究で挙げた先行研究による手法との比較も行っていく予定である．しかしながら，投稿によっては精度が著しく低くなるような例もある．以下には，このような例を受け現行の手法に対する問題点を挙げる．

まず第一に考えられるのは，関連語の取得や関連度の指標となるエッジの重み付けにおいて同時にタグ付けされる割合を取得しているが，この結果がリアルタイムに変動しうることである．この点は流行の変化などの現実的な問題に対応できる可能性がある一方で，手法の精度の測定においても結果が測定ごとに変動してしまう可能性がある．従って，これらの数値の取得方法には改善の余地があり，またそれらを使ってどのようにスコアを計算するかという手法についても様々な可能性が残されている．

また，前述の通り，Instagram は全世界で利用されている SNS であるため，日本語の投稿と英語の投稿の件数を比較すると，当然ながら英語の投稿のほうが圧倒的に多く，タグが使われる回数もその分多くなる．それゆえ，同じ言葉であっても，日本語で書かれたものよりも英語で書かれたもののほうが件数が多くなることも問題である．この問題は，提案手法において，言語の違いによる同義関係にある語同士を結んだ双方向グラフのエッジの重みが等しくならないという問題に直結する．

その他には，例えばデータ中に同一投稿者の投稿が多数含まれた場合に，その投稿者に特有のタグや，特有のタグの組み合わせがあった場合，現行の手法においてはそれらがスパムのような働きをすることがある．「同時にタグ付けされる割合」が高くなるためである．

Instagram 内だけで使われている造語や独特の表現についても同様である．例えば「Instagram における cafe」といった意味の「cafestagram」などの語が，造語にあたる．このような語はもとの語（この場合は「cafe」）と同時にタグ付けされることがほとんどであることから，提案手法では「cafe」のスコアを大幅に高めてしまう．

ここに挙げた問題以外にも，改善の余地は非常に多く残されていると考えており，今後は手法の性能改善に努めていきたい．

謝辞　本研究は JSPS 科研費 2628011, 26540163 の助成を受けたものです．

文　献

[1] Roi Blanco and Christina Lioma. Graph-based term weighting for information retrieval. *Inf. Retr.*, Vol. 15, No. 1, pp. 54–92, 2012.

[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] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pp. 1031–1040, New York, NY, USA, 2011. ACM.

[5] 大澤幸生, ベンソン ネルス E, 谷内田正彦. Keygraph : 語の共起グラフの分割・統合によるキーワード抽出. 電子情報通信学会論文誌. D-I, 情報・システム, I-コンピュータ, Vol. 82, No. 2, pp. 391–400, feb 1999.

[6] 小倉由佳里, 小林一郎. 単語の共起グラフを用いた重要文抽出による文書分類. 人工知能学会全国大会論文集, Vol. 27, pp. 1–4, 2013.

[7] 野本忠司. 確率モデルによる主題の自動抽出. 情報処理学会研究報告自然言語処理 (NL), Vol. 1995, No. 69, pp. 1–6, jul 1995.

[8] 野本忠司, 松本裕治. テキスト構造を利用した主題の推定について. 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, Vol. 96, No. 157, pp. 47–54, jul 1996.