Pixivにおけるキャラクタータグを検出する手法の検討

竹渕 瑛一[1] 山田 泰宏[1] 猪狩 知也[1] 鈴木 浩[1] 服部 哲[1] 速水 治夫[1]

概要：イラスト投稿型 SNS の一つである Pixiv では，多くの二次創作イラストが投稿されている．Pixiv では，ユーザがイラストに対し自由にタグ付けできる．また，二次創作イラストが多いため，タグの特徴として作品タイトルを表すジャンルタグと，そのキャラクターを表すキャラクタータグがよく付けられる．一方で，Pixiv ではこの 2 種類のタグについて区別を行なっていない．本研究では，与えられたイラストのタグ群からキャラクタータグを検出する手法（キャラクタータグ分類法）について検討を行った．提案手法は与えられたタグ群のジャンルタグに対して，よく付けられやすいタグをキャラクタータグとする．実験では提案手法の適合率と再現率，F 尺度について求め，評価を行った．

An Examination into Method of Character Tag Classification for Pixiv.

Eiichi Takebuchi[1] Yasuhiro Yamada[1] Tomoya Igari[1] Hiroshi Suzuki[1] Akira Hattori[1]
Haruo Hayami[1]

1. はじめに

近年，ユーザが自由にタグ付けのできるサービスが人気を集めている．その中でも Pixiv[3] が有名である．Pixiv ではユーザが自由にタグ付けをすることができる (図 1)．タグが付けられることで，タグを利用した検索を行うことが可能となる．

Pixiv では自由にユーザがタグ付けできるが，それぞれのタグがどのような意味を持っているかまでは設定できない．例えば，ある作品名を関するタグが付けられた時，望ましい動作としてはそのタグが作品名を表す，ジャンルタグであることを設定できる機能である．この機能がない場合，ユーザからはどのタグが作品名を表すか理解できても，システムからは理解することができない．

著者らは過去の研究 [1] により，Pixiv を対象とした二次創作のイラストに含まれているタグ群の中から漫画や映画，アニメ等の作品のタイトルとなるタグ（以下，ジャンルタグ [4]）を検出する手法について研究を行った．

タグ群の中からジャンルタグが区別されることによって，どのタグが作品名を表しているのかがシステムにも理解できるようになる．また，これを応用すれば，協調フィ

図 1 Pixiv のイラスト閲覧画面
投稿者: サッカン
イラスト ID: 31455155

[1] 神奈川工科大学大学院
KAIT, 1030, Shimo-Ogino, Atsugi, Kanagawa, Japan

ルタリング等により，似ているジャンル等を検索の候補に挙げることも可能となる．

一方で，二次創作のイラストのタグ群に含まれている特徴的なタグはジャンルタグだけではない．数多くの漫画や映画，アニメ等の作品には必ずキャラクターが登場し，オリジナルや二次創作を問わず，イラストの大半にはキャラクターが描かれている．従って，ジャンルタグがそのイラストのタグ群に存在するのならば，キャラクターとなるタグ（以下，キャラクタータグ）も存在していると考えられる．

キャラクタータグもジャンルタグと同様に，タグ群の中から区別されることにより，検索システムの利便性の向上を図ることが可能となる．特に，二次創作のイラストでは，描かれているキャラクターを知っていてもその名前を知らないことが多くある．2 ちゃんねるではイラストに描かれているキャラクターを質問するための専用のスレッド [5] や，過去に質問されたイラストから作者名とキャラクター名を検索するシステム [6] が存在している．

本研究では，Pixiv のイラストに含まれているタグから，あるジャンルにおいてキャラクターとなるタグを検出する手法について検討を行った．本論文では，その手法と実験的評価について述べる．

1 章では研究の概要，2 章では研究対象の現状と問題点，3 章では関連研究と提案手法の応用分野，4 章では提案手法の概要及び定式化，5 章では実験による提案手法の評価，6 章では提案手法の性能及び適合漏れに関する考察，7 章で本論文についてまとめる．

2. 研究の現状と問題点

Pixiv とは，イラスト投稿型 SNS の一つである．投稿したイラストを中心に，コメントやタグ付けを行うことでコミュニケーションを取る点がサービスの特徴である．

Pixiv では自由にタグ付けができる一方，システムではタグ群の中でどのタグがジャンルタグかは理解していない．描かれているジャンルをユーザが理解していても，システムはどのジャンルが描かれているのか理解することができない現状がある．

例えば，「ドラゴンボール」というタグは漫画もしくはアニメ作品のタイトルを表すタグである．コンテンツにはいくつかのタグが付けられているが，ユーザがジャンルタグを探すのは容易である．これは，ユーザが知識として「ドラゴンボール」という作品を知っていて，かつジャンルであると理解しているからである．

システムも同様に，「ドラゴンボール」というタグにジャンルを表すタグであることが情報として入力されていれば，タグごとにジャンルタグかどうか判定することでジャンルタグを検出することができる．しかし，多くのサービスではジャンルを表す情報は入力されていない．特に，国内のサービスでは顕著である．

そこで，検索システムの利便性を向上させるために，数多くのタグの中からジャンルタグを区別することを考える．しかし，大量のタグの中からジャンルタグを人の手で区別するには限界がある．大量の中からジャンルタグを区別するのに，自動的に処理してくれることが望ましい．

著者らは Pixiv を対象に，タグとタグの関連からジャンルタグを検出する研究を行った [1]．Pixiv のタグにはジャンルタグとなる一定の傾向が認められた．この研究のアルゴリズムを利用することで，タグ群からジャンルタグを検出することができる．

ジャンルタグが区別されることにより，もしユーザが閲覧しているイラストのジャンルがわからなかったとしても，どのタグがジャンルタグであるか提示することが可能となる．他にも，よく閲覧されるジャンルやホットなジャンルを提示したり，協調フィルタリングを応用することで，あるジャンルに対して他のユーザがよく見るジャンルを提示することが可能となる．

一方で，あらゆる作品には登場人物（以下，キャラクター）が存在する．キャラクターの多くには名前があり，Pixiv でもその名前を冠するキャラクタータグが付けられる．例えば，「ドラゴンボール」であれば，「孫悟空」や「ヤムチャ」，「フリーザ」などのキャラクターが該当する．

このようなキャラクタータグも無数に存在している．また，その数はジャンルタグよりも多いことが予想され，キャラクタータグであるか否かを区別するのは困難を極める．これも著者らの過去の研究のように，タグ群の中から自動的に検出できることが望ましい．

そこで著者らはキャラクタータグを見つけ出すアルゴリズムについて研究を行った．従来の検索システムでキャラクターを基準とした検索を行うには，ユーザがキャラクターの名前を知っている必要があった．提案手法によりキャラクタータグを区別することで，ユーザがキャラクターの名前を知っていなくとも検索を行うことが可能となる．

キャラクタータグが区別されることにより，もしユーザが閲覧しているイラストのキャラクターがわからなかったとしても，それが誰なのかタグとして提示することが可能となる．また，ジャンルごとに人気のキャラクターや，最近閲覧されやすいキャラクター，あるキャラクターと似ているキャラクター等の提示も可能となる．

3. 関連研究及び本研究の応用分野

キャラクタータグが区別されることによって，様々な研究用途で利用することが可能となる．

例えば，画像検索である．関連研究として，ぱろすけによる大規模 AV 画像データベースと類似顔画像検索を用いた AV 検索システムがある [2]．DMM.com に登録されている静止画像と出演している女優が紐付けられているた

め，事前に女優ごとに統計的顔画像データを作成することで，顔認識による類似女優の検索ができることが特徴である (図 2).

図 2 ぱろすけによる顔画像をもとに似た顔の人が出ている AV を検索するツール

本研究も無数に存在するタグの中からキャラクターとなるタグを区別するため，イラストとキャラクタータグが紐付けられることとなる．従って，ぱろすけによる研究をキャラクターの顔画像に応用することで，与えられた画像から似ているキャラクターを検索することが可能になると考えられる．

また，応用例の一つとして，キャラクターごとに付けられたタグの傾向から，別のジャンルにおける似ているキャラクターを検索することも考えられる．ぱろすけによる研究では顔認識をベースとしていた．タグは主にキャラクターの特徴等が付けられる傾向がある．これにより，似たようなタグからキャラクターの特徴が似通っているキャラクターを推薦することができると考えられる．

ぱろすけによる研究では，顔画像を利用していることが特徴である一方，その画像に関するメタ情報等は考慮されない．二つの応用例を組み合わせることにより，メタ情報を考慮したキャラクター顔画像検索システムを構築することができると考えられる．

4. キャラクタータグ分類法

本章では提案手法の概要及び定式化について述べる．

4.1 提案手法の概要

提案手法はジャンルタグとキャラクタータグが親子関係にあることに着目した．

ジャンルタグは作品名を示すタグであり，キャラクタータグはその作品の登場人物を表すタグである．従って，キャラクタータグが付けられた件数はジャンルタグを付けられた件数を上回ることは考えられない．

また，キャラクタータグはジャンルタグが確実に付けられなければ，どのような作品に登場するキャラクターなのか判別することが難しい．これは，キャラクタータグで検索を行った時にジャンルタグが極めて高い確立で共起しなければ，親より子の方が多く共起することとなり，階層構造的に自然ではない．

これらのことから，キャラクタータグとは，ジャンルタグよりもタグ付けされた件数が少なく，キャラクタータグで検索を行った時，ジャンルタグが極めて高い確率で共起するタグであると考えられる．

4.2 提案手法の定式化

前節より，提案手法の定式化を行う．

あるタグを t としたとき，イラストに含まれているタグ群を $T = \{t|t_1, ..., t_n, 1 \leq n \leq 10\}$ とする．ここでは便宜的にジャンルタグを j，キャラクタータグを c で表す．従って，タグ群に含まれるジャンルタグは T_j，キャラクタータグを T_c と表される．タグの付けられたイラストの件数をタグの絶対値 $|t|$ と表す．

ある任意のタグ x で検索を行い，x を含むタグ群の集合を得る関数を $R(x) = \{T|T_1, ..., T_m, x \in T\}$ とする．

与えられたタグ群 T からキャラクタータグを得る関数を $f(T)$ とする．前節で述べたキャラクタータグの規則を表すと，下式のようになる．

$$
\begin{aligned}
&f : T \to c \\
&f : T \mapsto c = \{c| \max_{t \in T} P(T_j|t), 1 \leq |c| \leq |T_j|, T_j \in T\}
\end{aligned}
\tag{1}
$$

ただし，本研究では複数作品のコラボレーションを行なっているイラストのタグ群については考慮しないため，T_j は可能な限り 1 つである必要があり，$0 < |T_j| < 2$ でない場合，ジャンルタグが誤検出される可能性が極めて高い．また，提案手法は 1 件のキャラクタータグのみ検出することができる．複数件キャラクタータグが含まれるタグ群に関しては，最も $P(T_j|t)$ の値の高いタグが c の候補となる．

式 2 は，タグ群のジャンルタグ T_j で検索を行い，それらのタグ群から任意のタグ x が含まれる共起確率を表している．

$$
P(j|x) = \frac{|R(j) \cap x|}{|R(j)|} \tag{2}
$$

提案手法はタグ群の各タグが，ジャンルタグで検索した場合の共起確率について求め，最もジャンルタグに対して付けられやすいタグをキャラクタータグとしている．

5. 提案手法の実験的評価

本研究では，試作システムに提案手法を実装した．試作システムを利用して提案手法の実験的評価を行い，有用性について確認した．本章では実験の方法及び実験結果について述べる．

5.1 実験の概要

本研究では 2 種類の実験を行った．

実験 1 ジャンルタグを対象とした再現率を求める

実験 2 キャラクタータグを対象とした適合率を求める

実験 1 は対象のジャンルタグを含むタグ群に対して提案手法を適用したとき，キャラクタータグが適合するか調査を行った．実験 1 は提案手法の正確性を求めることを目的とし，対象のジャンルタグにおけるキャラクタータグが全て適合するか調査を行う．

実験 2 はキャラクタータグを含むタグ群に対して提案手法を適用した時，キャラクタータグが適合するか調査を行った．実験 2 は提案手法の網羅性を求めることを目的とし，対象のキャラクタータグが全て適合するか調査を行う．

5.2 実験環境

試作システムによる実験を行うためにデータベースを作成した．データベースは Pixiv に対してクローリングを行い，ユーザ情報（ユーザ ID，投稿したイラスト ID）及びイラスト情報（イラスト ID，タグ）をデータベースに格納した．

クローリングはユーザ ID をもとに行った．2013 年 1 月 28 日に現存するユーザ ID の若い順番から 124,944 件取得し，それらユーザの投稿したイラストを 1,040,104 件を取得した．なお，取得したイラストに付けられたタグは 572,933 種類あった．

5.3 実験方法

本節では各実験における実験方法について述べる．

5.3.1 実験 1

実験 1 では，「けいおん!」，「とある科学の超電磁砲」，「らき☆すた」の 3 種類のジャンルタグについて実験を行った．実験はそれぞれのジャンルタグを含むタグ群を全件取得し，提案手法を適用した．それぞれのジャンルタグにおいて，登場するキャラクタータグが提案手法の候補になったことで適合されたとする．

実験 1 は提案手法の正確性を調べるための実験である．実験で扱うジャンルタグは，可能な限りキャラクターが少なく，投稿数の多いタグを選んだ．キャラクターが多い場合，適合するキャラクタータグを網羅することが難しくなるからである．

5.3.2 実験 2

実験 2 では，「御坂美琴」，「綾波レイ」，「如月千早」の 3 種類のキャラクタータグについて実験を行った．実験はそれぞれのキャラクタータグを含むタグ群を全件取得し，提案手法を適用した．それぞれのキャラクタータグが提案手法の候補になったことで適合されたとする．

実験 2 は提案手法の網羅性を調べるための実験である．実験で扱うキャラクタータグは，投稿数が多いタグを選んだ．また，提案手法は複数作品とのコラボレーションを行なっているイラストのタグ群については考慮していないため，可能な限り複数のジャンルタグを付けられにくいキャラクターを選んだ．

なお，「御坂美琴」のジャンルタグは「とある科学の超電磁砲」であり，「綾波レイ」は「エヴァンゲリオン」，「如月千早」は「アイドルマスター」がジャンルタグであると想定している．表記ゆれについては考慮しない．

5.4 実験結果

本節では前節の方法で行った実験の結果について述べる．

5.4.1 実験 1

実験 1 では，「けいおん!」，「とある科学の超電磁砲」，「らき☆すた」を含むジャンルタグを対象として，それらを含むタグ群に対して提案手法を適用し，キャラクタータグの再現率を求めた．ただし，表記ゆれに関しては積極的に適合したものとする．また，カップリング等による複数キャラクタータグを省略する表記に関しては適合しないものとする．

表 1 実験 1 の結果

tag	genre	adapt	recall(%)
けいおん!	5341	4519	84.6
とある科学の超電磁砲	1121	997	88.9
らき☆すた	4282	3561	83.1

genre は対象となるジャンルタグを含んだタグ群の数であり，adapt は適合した件数，recall は再現率である．

実験より，平均再現率は 85.3%となった．表記ゆれも含め，ほぼ確実にキャラクタータグが適合していることがわかった．

5.4.2 実験 2

実験 2 は，「御坂美琴」，綾波レイ」，「如月千早」を含むキャラクタータグを対象として，それらを含むタグ群に対して提案手法を適用し，キャラクタータグの適合率を求めた．

表 2 の subject は関数 $f(T)$ の規則に適合するタグの件数であり，genre は対象のジャンルタグの総数，adapt は適合したキャラクタータグの数，precision は適合率である．

実験より，全てのキャラクタータグの適合率が 100.0%となった．提案手法は特定のキャラクタータグに対しては，

表 2 実験 2 の結果

tag	subject	genre	adapt	precision(%)
御坂美琴	621	12512	621	100.0
綾波レイ	628	1179	628	100.0
如月千早	1003	18358	1003	100.0

確実にタグ群から検出できることがわかった.

6. 考察

本章では実験の結果より，提案手法の性能と適合漏れの傾向について述べる.

6.1 提案手法の性能

前章の実験より，平均再現率は 85.3%，平均適合率は 100.0%となった.

提案手法はタグ群の中でジャンルタグが決まっていればほぼ確実にキャラクタータグを検出することが可能であることがわかった．特に，キャラクタータグを含んでいるタグ群を対象とした実験 2 では，確実に対象のキャラクタータグが検出されていた.

平均再現率及び平均適合率から F 尺度を求めると，0.926 となった．F 尺度は適合率と再現率の調和平均によって求められる指標であり，正確性と網羅性の総合的な評価に使われる．F 尺度は下式によって求められる.

$$F - measure = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{3}$$

実験により適合したタグが 8 割を超えていることから，提案手法とキャラクタータグに関して明確な法則性が存在しているものと考えられる．機会学習により頻繁に検出されるタグを学習することにより，より正確にキャラクタータグを検出することも可能であると考えられる.

6.2 実験 1 での適合漏れ

実験 1 の適合漏れのタグ群に関しては，キャラクタータグが含まれていないケースが多かった．ジャンルタグのみタグ付けを行なっているイラストの多くは，他に付けられているタグが少ない傾向にあることがわかった．また，イラストのテーマが他のジャンルタグであり，ついでに対象のジャンルタグが付けられているような場合は，タグのほとんどは他のジャンルタグに付けられやすい傾向があった.

また，提案手法の適合漏れの代表例として，ジャンルタグの表記ゆれが挙げられる．仮にキャラクタータグが含まれていたとしても，キャラクタータグ以外のタグがジャンルタグに対してよく付けられている場合として，最もジャンルタグの表記ゆれが多かった．「らき☆すた」では「らきすた」のように省略していたり，「けいおん!」では「けいおん」,「けいおん!!」,「けいおん!3 年 2 組」等，さまざまなジャンルタグの表記ゆれが見られた．「けいおん!」では，実験で用いたタグ群の総数に対してジャンルタグの表記ゆれの占める割合が 5.2%，適合漏れの中でジャンルタグの表記ゆれの占める割合が 33.8%にも上った.

提案手法がジャンルタグの表記ゆれに関しても積極的に検出してしまう傾向から，事前にジャンルタグの表記ゆれを提案手法で検出しておき，ジャンルタグの表記ゆれリストを作成しておくべきであると考えられる．再び提案手法を適用する場合，そのジャンルタグの表記ゆれリストに登録されているタグをあらかじめ除外しておくことで，適合漏れを防ぐことが可能なのではないかと考えられる.

上記のような工夫を施すことで，平均再現率が約 90%，F 尺度が約 0.95 まで改善することが可能ではないのかと考えられる.

7. おわりに

本論文では，Pixiv のイラストに含まれるタグ群から，キャラクタータグを検出する手法について述べた．提案手法は，ジャンルタグを含むタグ群の中で，ジャンルタグと共起しやすいタグをキャラクタータグとする手法である.

提案手法の評価のため実験を行ったところ，平均再現率は 85.3%，平均適合率は 100.0%，F 尺度が 0.926 となった．結果としては概ね満足できるレベルであり，機会学習を利用することでほぼ確実にキャラクタータグを検出することが可能になると考えられる.

キャラクタータグを検出できたことにより，キャラクターに付けられやすいタグの調査をすることが可能となった．今後の展望として，キャラクターの特徴を示すタグを与えることで，その特徴を持ったキャラクターを推薦することのできるシステムを研究したい.

参考文献

[1] 竹渕瑛一: Pixiv の二次創作イラストに含まれるジャンルタグの自動分類 研究報告グループウェアとネットワークサービス（GN）, Vol.86, No.24, pp1-5 (2013).

[2] ぱろすけ: 大規模 AV 画像データベースと類似顔画像検索を用いた AV 検索システム, あの人の研究論文集, Vol.3, No.2, pp1-4 (オンライン), 入手先〈http://www3.kitanet.ne.jp/ narumin/anohito_CFP.htm〉 (2012).

[3] ピクシブ株式会社: Pixiv, 入手先〈http://www.pixiv.net/〉 (参照 2013-05-15).

[4] ピクシブ株式会社: ジャンル, ピクシブ百科事典 (オンライン), 入手先〈http://dic.pixiv.net/a/ジャンル〉 (参照 2013-05-15).

[5] 不明: 【この娘誰?】気楽に詳細を聞いてみるスレッド 279, 入手先〈http://pele.bbspink.com/test/read.cgi/ascii2d/1366760063/l50〉 (参照 2013-05-15).

[6] ascii2d: 二次元画像詳細検索, 入手先〈http://www.ascii2d.net/imagesearch〉 (参照 2013-05-15).