Vol.2013-GN-86 No.24
Vol.2013-CDS-6 No.24
2013/1/17

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

竹渕 瑛一[1]　鈴木 浩[2]　服部 哲[3]　速水 治夫[3]

概要：近年，投稿されたコンテンツに対して利用者が自由にタグを付けられるサービスが増加している．それに伴い，コンテンツに付けられたタグ群を自動分類する研究も盛んになっている．自動分類の一つとしてタグの階層化が挙げられる．タグの親子関係を構築することにより，検索の利便性を向上させる研究である．一方で，現状の研究では不特定のタグを階層化するため，目的のタグが見つからない問題がある．本論文は Pixiv を対象に，二次創作イラストに付けられるジャンルタグを自動分類する手法（ジャンルタグ分類法）について述べる．ジャンルとは，2 次創作における原作を意味している．ジャンルタグ分類法とは，対象のイラストのタグ群から，対象のイラストと同様なタグを付けたイラストとの相互関係によりジャンルタグの推定を行う手法である．ジャンルタグの分類を行うことによって，階層化されたタグのうち，どの階層にジャンルタグが存在するか特定できるようになる．

キーワード：ジャンルタグ, Pixiv, フォークソノミー, 自動分類, データマイニング, 集合知

1. はじめに

近年，フォークソノミーにおけるタグの自動分類が研究分野として注目を浴びている．例えばタグの階層化である．タグの階層化とは，利用者が自由にタグを付けることのできるフォークソノミーのように，階層構造を意識せずに付けられたタグ群に対して，コンテンツ同士のタグの相関関係から自動的に階層構造を推定する手法である．

タグの階層化は無数に存在するタグを整理し，コンテンツに自動的なタグ付けを行う研究 [1] もあるため，研究分野としての応用範囲は広いと考えられる．タグの階層化は無数に存在するタグを整理し，検索システムの利便性を向上させる点では有効である．一方で，無数のタグを階層化するため，検索結果として得たいタグがすぐに見つからない場合もある．特に，どの階層にジャンル（2 次創作における原作）が存在するかわからない問題がある．

著者らは，Pixiv[3] を対象にジャンルタグの自動分類の手法（ジャンルタグ分類法）について研究を行った．ジャンルタグとは，Pixiv における二次創作イラストの原作を表すタグを指している．

ジャンルタグ分類法は対象のイラストのタグ群から，対象のイラストと同様なタグを付けたイラストとの相関関係によりジャンルタグの推定を行う．ジャンルタグ分類法は無数に存在するタグの中からジャンルを意味するタグを分類することによって，今まで不足していたジャンルに関する情報をタグに与えることができる．これにより，タグの階層化において不足していたタグの情報が補完され，検索システムのさらなる利便性の向上を図ることが可能になる．

本論文では提案手法であるジャンルタグ分類法とその評価について述べる．1 章では研究の概要と背景について述べる．2 章ではタグの自動分類に関連する研究と現状について述べる．3 章では提案手法であるジャンルタグ分類法のアルゴリズムと定式化について述べる．4 章ではジャンルタグ分類法を適用した分類評価実験とその結果について述べる．5 章では 4 章の考察を行い，6 章で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

タグの階層化についてはニコニコ動画やソーシャルブックマークを中心に広く研究が行われている．その中でもニコニコ動画に関連する研究として，相川勇気らによる研究 [4] と伊藤栄典らによる研究 [5] が挙げられる．

伊藤栄典らの研究では，投稿された動画のタグを ISR 手法によってタグの階層化を行なっている．単語 u の文書頻度が単語 v の文書頻度を上回り，なおかつ単語 v が出現する中で単語 u の共起確率が α を超えた場合，単語 u は単語

[1] 神奈川工科大学大学院博士前期課程
[2] 神奈川工科大学大学院博士後期課程
[3] 神奈川工科大学

ⓒ 2013 Information Processing Society of Japan

v の親であるという．α の値を高く設定することで単語同士の適合率を高くすることも可能であるが，低くした場合はノイズが発生するとしている．

相川勇気らの研究では，ブラウザの視聴履歴からニコニコ動画の視聴履歴を取得し，その視聴履歴をタグで分類するための試作システムを実装している．タグ a を検索対象とした時，タグ a の含まれる動画を一覧表示にする．また，タグ a の含まれる動画を AND 検索でタグを取得し，取得したタグとその登録件数をプルダウンリストで管理している．これにより，タグ a が付けられ，なおかつタグ b，タグ c のように，階層構造的に検索結果を辿ることができる．

これらの研究は投稿された動画のタグの階層化を行っている．伊藤栄典らの研究は無数に存在するタグの階層化を行うことで検索システムの利便性向上を目的とし，類似動画の推薦に有効であると指摘している．相川勇気らの研究はブラウザの視聴履歴からタグを階層構造的に AND 検索をかけることで，視聴した動画を再び閲覧するときに有効であると指摘している．

伊藤栄典らと相川勇気らの研究は，目的の動画の検索や面白い動画の検索に対して有効である．しかし，これらの研究は無数のタグを階層化するため，利用者が階層構造を辿ろうとする場合においてはまだ議論の余地がある．

例えば，利用者が検索キーワードを忘れてしまった場合である．利用者はあるジャンルのイラストの検索を行いたいと考えるが，そのジャンルタグの名前を忘れてしまっている場合，既存の階層化されたタグのみでは検索結果を得ることができない．一方で，利用者は忘れてしまったジャンルに付けられる特徴的なタグを幾つか把握している．このような利用者のケースでは，利用者は忘れてしまったジャンルタグをすぐに推薦して欲しいと考えるのに対し，タグの階層化だけでは利用者の目的を達成することができない．

本研究では，無数のタグのうちどのようなタグがジャンルタグであるか分類するものである．本研究では，上記のような利用者のケースに対して有効である．タグの一つ一つがジャンルタグであるかどうかを表すことができるようになるため，検索時の検索結果としてジャンルのみを提示することが可能となる．

3. ジャンルタグ分類法

ジャンルタグ分類法は対象となるイラストのタグ群からジャンルとなるタグを確率的に分類する手法である．本章ではジャンルタグ分類法のアルゴリズムとその定式化について述べる．

3.1 ジャンルタグ分類法のアルゴリズム

まず，分類の対象となるイラストのタグ群から任意のタグを選び出す．任意のタグ x から同様のタグを含むイラストについて検索を行い，この検索によって得られたタグ群の集合をページと呼ぶ．

分類の対象となるイラストのタグ群からさらにもう一つ任意のタグ y を選び出し，ページ内のタグ群の集合にその任意のタグが含まれているかどうか調べる．この時，ページ内に存在した任意のタグ y が含まれているタグ群の総数を，任意のタグ x で検索して得られたイラストの件数で割った数が，任意のタグ x でページを取得し，さらに取得したページから任意のタグ y が含まれる共起確率となっている．

これを分類の対象となるイラストに含まれる全てのタグについて繰り返す．これを元にクロス集計表を作成する．例として**表 1** を挙げる．表 1 は任意のタグ x を行とし，任意のタグ y を列とする．

表 1 クロス集計表の例
Table 1 An example of cross summary sheet

	A	B	C	D
A		0.0	0.0	0.8
B	1.0		0.0	1.0
C	0.0	0.0		0.0
D	0.2	0.0	0.0	

表 1 では，分類の対象となるイラストのタグを A, B, C, D としている．例えば，ページを取得する任意のタグを A とした場合，ページ内に B が含まれる共起確率は 1.0 となる．

ジャンルタグの推定は列ごとに共起確率をスコアとして総和を取り，最もその数値が高い列，すなわち任意のタグ x がジャンルタグである．表 1 では，A 列が 1.2，B 列が 0.0，C 列が 0.0，D 列が 1.8 となっているため，タグ D がジャンルタグであると推測される．

3.2 アルゴリズムの定式化

あるイラストにおけるタグ群を $T = \{t_0, t_1, ..., t_n\}$ とする．分類の対象となるイラストのタグ群は $\hat{T}$ と表す．対象のタグ群 $\hat{T}$ の中から任意のタグ t で検索して得られたイラストのタグ群をページ $P_t = \{T_0, T_1, ..., T_m\}$ とする．

検索対象のタグ x で検索したページ内にタグ y が現れる共起確率を $f(y|x)$ とする．

$$f(y|x) = \frac{|P_x \cap y|}{|P_x|} \tag{1}$$

さらに，分類の対象となるタグ群を引数とする関数 $J(T)$ により，それぞれのタグについて式 1 による演算を行う．

ⓒ 2013 Information Processing Society of Japan

このとき最大値を示す t がジャンルタグである．

$$J(T) = \max_{t \in T} \sum_{i=0}^{n} f(T_i|t) \quad (2)$$

4. 分類評価実験

3 章で述べたアルゴリズムに基づき，ジャンルタグ分類法の分類評価実験を行った．本章では分類評価実験の結果と考察について述べる．

4.1 特定のジャンルタグの適合率

特定のジャンルタグの適合率について，「涼宮ハルヒの憂鬱」「らき☆すた」「東方」「VOCALOID」の 4 種類のタグを対象に実験を行った．それぞれのタグを含むイラストを検索し，取得件数だけ分類が成功しているか調べている．

表 2 はその結果である．

表 2 特定のジャンルタグの適合率（%）
Table 2 Apply the Genre Tag Classification Method by a genre tag for a success rate.

タグ名＼m	取得数	10	20	平均
涼宮ハルヒの憂鬱	2733	60.8	71.6	63.4
らき☆すた	4329	91.2	90.8	91.0
東方	110597	98.5	98.9	98.5
VOCALOID	16591	59.8	42.7	52.4

表 2 における行の数値はページにおけるイラスト取得件数 m を表している．

この実験では，ジャンルタグが良好な結果を得られる場合とそうでない場合に分かれた．分類に失敗する例として，対象のイラストにタグが 2 つしか付けられていない，キャラクターが 1 つのジャンルとして成立している，ジャンルを持っていないキャラクターがいる，作品内に特別なジャンルが存在する，1 つのイラストに複数ジャンルが設定されているなどが挙げられる．

対象のイラストにタグが 2 つしか付けられていない場合では，ジャンルタグとキャラクタータグの組み合わせならばジャンルタグの分類は可能であるが，一方でジャンルタグが設定されていたとしても，もう一方が作品情報タグとは無関係なタグが設定されていれば誤分類の原因になる．

キャラクターが 1 つのジャンルとして成立している例では，「初音ミク」が誤分類を起こしやすい例として挙げられる．「VOCALOID」は「初音ミク」や「鏡音リン」のキャラクターを含んでいるが，「VOCALOID」のタグはあまり付けられていない．表 2 の平均を見ても，およそ半分のイラストが誤分類されているのがわかる．VOCALOID はジャンルではなく，キャラクター群もしくはジャンルとして捉える見方がある．このことがジャンルタグとして「VOCALOID」が付けられない理由なのではないかと考えられる．

ジャンルを持っていないキャラクターとしては，「備長たん」や「ひこにゃん」などのインターネットコミュニティ発祥のキャラクターやご当地キャラクターのことを指している．これらはジャンルを持っていないため，他の登録件数の多いタグやそのキャラクターの特徴を表しやすいタグを誤分類の傾向にあった．

1 つのイラストに複数ジャンルが設定されている場合として，例えば「涼宮ハルヒの憂鬱」と「らき☆すた」のタグが同時に存在する場合が挙げられる．「らき☆すた」では「涼宮ハルヒの憂鬱」のパロディが少なからず存在している．本来であれば「涼宮ハルヒの憂鬱」がジャンルタグの候補として挙げられるべきであるが，「らき☆すた」のほうがタグの登録件数が多い．これが誤分類の原因となっている．

4.2 無作為抽出におけるジャンルタグの再現率

1000 件のイラストを無作為に抽出し，抽出したイラストのタグ群からジャンルタグの分類を行った時の再現率について実験を行った．**表 3** はその結果である．

表 3 無作為抽出における再現率（%）
Table 3 Apply the Genre Tag Classification Method to sampling from the randomize for a success rate.

項目＼m	10	20	平均
ジャンルタグを除く	56.0	53.3	56.1
ジャンルタグを含む	64.9	65.1	66.9

各イラストにおける分類の成功，失敗に関しては検索エンジンなどで調査を行なった上で計上している．なお，行の数値はページにおけるイラスト取得件数 m を表している．

無作為の場合における再現率はジャンルタグを除いておよそ 56.1%，ジャンルタグを含めて 66.9%で成功することがこの実験でわかった．

誤分類の多くは「初音ミク」「オリジナル」などのようなジャンルに近いタグや，「漫画」「版権」「落書き」などのようなイラストの形態を示すタグ，「女の子」「ケモノ」「制服」などの描かれたキャラクターの特徴を示すタグが候補として挙げられる．

ジャンルに近いタグに関しては，それよりの上位のタグが存在しないか，もしくはその上位のタグの登録件数が少ない場合に誤分類される傾向が見られる．例えば，「初音ミク」の登録件数は 268,695 件であるのに対し，「初音ミク」と「VOCALOID」が同時に登録されているタグの登

ⓒ 2013 Information Processing Society of Japan

録件数は 138,437 件である（2012 年 11 月 26 日時点）.「初音ミク」のみのほうが本来のジャンルタグとの組み合わせよりも多く登録されている.

「漫画」「版権」「落書き」などのようなイラストの形態を表すタグは，その登録件数が多いことと，ジャンルタグが設定されていない場合などで特に多く誤分類される傾向があった．特にこれは「オリジナル」が設定されるべきであるイラストに多く見られた.

「女の子」「ケモノ」「制服」などの描かれたキャラクターの特徴を示すタグは，検索時における登録件数が多いことにより，誤分類される結果となった.

4.3 ジャンルタグ分類法によるオリジナルタグ，ジャンルタグの適合率

ジャンルタグ及びキャラクターやイラストの特徴を示すタグが含まれたタグ群に対してジャンルタグ分類法を適用し，オリジナルタグの適合率について実験を行った．この実験では，ジャンルタグ分類法を適用した結果から，オリジナルタグ及びジャンルタグが含まれている割合を求め，タグごとにオリジナルタグとジャンルタグの適合率について調査を行なっている.

表 4 ジャンルタグにおける適用結果（%）
Table 4 A result of Genre Tag Classification Method from genre tags.

タグ＼割合	取得数	誤分類	original	genre
東方	1000	1.2	0.0	0.0
らき☆すた	988	11.6	0.2	1.6
涼宮ハルヒの憂鬱	969	42.6	0.1	2.4
VOCALOID	995	33.6	0.0	1.8

表 4 は，ジャンルタグに対してジャンルタグ分類法を適用し，取得数からオリジナルタグ及びジャンルタグの適合率を求めている.

表 2 の結果と同様に，「涼宮ハルヒの憂鬱」及び「VOCALOID」のタグでは多くのタグが誤分類される結果となった．しかし，オリジナルタグや対象のジャンルタグ以外のジャンルタグの適合率は総じて低い結果となった．この結果により，ジャンルタグにはオリジナルタグやその他のジャンルタグが含まれにくい傾向があることがわかった．また，ジャンルタグから誤分類されるタグの多くは，オリジナルタグやジャンルタグ以外のタグが多く含まれることもこの表からわかった.

表 5 は，表 2 及び表 3 で述べたようなキャラクターやイラストの特徴を示すタグに対してジャンルタグ分類法を適用し，取得数からオリジナルタグ及びジャンルタグの割合を求めている.

最も誤分類の少ないタグはオリジナルタグである．オリジナルタグが誤分類された場合，オリジナルタグは表さないため，0.0%である．ジャンルタグが分類された場合については 0.1%と低く，オリジナルタグはジャンルタグが含まれにくいと考えられる.

表 5 特徴的なタグにおける適用結果（%）
Table 5 A result of Genre Tag Classification Method from characteristic tags.

タグ＼割合	取得数	誤分類	original	genre
落書き	882	89.0	16.5	27.8
女の子	966	59.9	41.6	9.7
制服	936	92.9	39.6	7.8
オリジナル	956	8.3	0.0	0.1

「落書き」「女の子」「制服」などのイラストやキャラクターの特徴を示すタグについては，表 4 と比べると誤分類される割合が極めて多いことがわかる．これはキャラクターやイラストの特徴を示すタグであるため，様々なジャンルタグのみならず，オリジナルタグにも付けられる傾向にあるからだと考えられる.

5. 考察

4.1 節及び 4.2 節の実験結果より，無作為抽出でのジャンルタグ分類法の適用においては再現率が低く，ジャンルタグ分類法単体で分類を行うだけではうまく分類されないことがわかった．特定のジャンルタグに対してジャンルタグ分類法を適用した場合も，ジャンルタグによって適合率がまばらである.

一方で，4.3 節の実験結果より，ジャンルタグ分類法によって分類されたタグから，分類されたタグを含むイラストに対してジャンルタグ分類法を適用することにより，分類されたタグ以外が分類される結果から，ジャンルタグはオリジナルタグの適合率が極めて小さいことがわかった．4.3 節の実験のように，分類されたジャンルタグから検索を行い，オリジナルタグの適合率を調べるような操作を再帰的に行うことで，ジャンルタグの分類のみならず，二次創作のイラストであるかオリジナルのイラストであるかも判別が可能になると考えられる.

これらの考察から再帰的操作を定式化する．$J(T)$ によって得られたタグを格納するためのタグ群を $R = \{r_0, r_1, ..., r_n\}$ と定義する．ただし，r_0 はジャンルタグ分類法を最初に適用したタグ $J(\hat{T})$ のことである．r_n で求められたタグによって取得したページ P_{r_n} のそれぞれのタグ群に対し，ジャンルタグ分類法を適用する．このときの集合を $Q_{r_n} = \{q_0, q_1, ..., q_n\}$ と定義する．これらの式の n は任意の定数である.

与えられたタグ群の中で最も共起確率の高いタグを選び出す関数を $\xi(T)$ とする.

ⓒ 2013 Information Processing Society of Japan

$$\xi(T) = \max_{t \in T} \frac{|t \cap T|}{|T|} \tag{3}$$

タグ群 R を求めるための漸化式を下式のように定義する．

$$\begin{aligned} r_0 &= J(\hat{T}) \\ r_{n+1} &= \xi(Q_{r_n}) \end{aligned} \tag{4}$$

このとき，タグ群 R の中で最も共起確率の高いタグが，$\hat{T}$ におけるジャンルタグであると考えられるため，最終的にジャンルタグは $\xi(R)$ により求めることができる．ただし，選び出されたタグの共起確率がある閾値以下の場合，$\hat{T}$ にはジャンルタグが含まれていないと考えられる．

4.3 節より，式 4 における共起確率の閾値は 50%より大きい数値が適当であると考えられる．これは，ジャンルタグであれば 4.3 節の表 4 より，繰り返し同様なジャンルタグが分類されることが想定できるからである．

ただし，このとき R の濃度が充分でないと分類に失敗することも考えられる．そこで，Q_{r_n} におけるオリジナルタグの適合率を求める関数 $O(r_n)$ を定義する．ここではオリジナルタグを o とする．

$$O(r_n) = \frac{|Q_{r_n} \cap o|}{Q_{r_n}} \tag{5}$$

式 5 より，オリジナルタグの適合率が 2〜5%以上の場合，濃度が十分ではないと考え，再度 r_{n+1} について計算を行う．ただし，Q_{r_n} の濃度よりオリジナルタグの適合率が 20%以上だった場合は，タグ群 $\hat{T}$ のイラストはオリジナルであると考え，計算を打ち切る．

式 3 から式 5 の操作を行うことで，より高い再現率でジャンルタグの分類を行うことが可能であると考えられる．

6. 本論文のまとめと今後の展望

本論文ではイラストに含まれているタグ群からジャンルタグを分類する手法（ジャンルタグ分類法）について述べた．分類評価実験より，ジャンルタグ分類法を特定のジャンルタグを含むイラスト及び無作為抽出によって選び出されたイラストに対して適用し，ジャンルタグ分類法における適合率と再現率について調査した．また，この 2 つの実験からジャンルタグ及びイラストの特徴を示すタグを含むイラストについてジャンルタグ分類法を適用し，オリジナルタグ及びジャンルタグの適合率について実験を行った．実験結果を踏まえ，考察において現状のジャンルタグ分類法について至らなかった点について考慮し，対象のイラストに含まれるタグ群だけでなく，分類されたタグを含むイラストについて再帰的にジャンルタグ分類法を適用することで，ジャンルタグ分類法の性能を向上させることが可能となった．

この研究により，階層化されたタグに対し，ジャンルタグ及びオリジナルタグの情報を付加することが可能となり，検索の利便性を向上させることができると考えられる．

今後の展望として，キャラクターを示すタグ及びイラストやキャラクターの特徴を示すタグの分類を行うことで，キャラクターを主体とした協調フィルタリングを行うための研究を行う予定である．

謝辞 本論文中の数式においては，神奈川工科大学の徳弘一路准教授にご教示を頂いたことに深謝する．神奈川工科大学大学院博士前期課程の相川勇気氏には，本研究のきっかけとなるアイディアを頂いたことに深謝する．

参考文献

[1] 風間淳一：教師なし隠れマルコフモデルを利用した最大エントロピータグ付けモデル，言語処理学会．自然言語処理，11(4)，2004-10，pp. 3-24
[2] ピクシブ百科事典 - ジャンル，http://dic.pixiv.net/a/%E3%82%B8%E3%83%A3%E3%83%B3%E3%83%AB
[3] Pixiv，http://www.pixiv.net/
[4] 相川勇気：動画のタグを視聴履歴の検索キーワードとして利用する動的多段絞り込み検索システム，情報処理学会．マルチメディア，分散，協調とモバイル（DICOMO2012）シンポジウム，p545-550
[5] 伊藤栄典：動画投稿サイトで付与された動画タグの階層化，情報処理学会研究報告．MPS，数理モデル化と問題解決研究報告 2010-MPS-81(17)，1-6，2010-12-09

ⓒ 2013 Information Processing Society of Japan