Vol.2013-ICS-170 No.3
2013/3/10

ニコニコ動画における共起関係を用いたタグの階層化

高橋 文彦[†1,a)]　山本 雅人[†2]　古川 正志[†2]

概要：Web 上では動画共有サービスの利用，特に国内では YouTube やニコニコ動画の利用が盛んである．ニコニコ動画ではタグにより動画が分類されている．一般に階層化されたタグをユーザに提示すると，ブラウジング効率が上がることが知られており，階層化手法も提案されている．本研究では，共起関係によるタグの類似度とネットワークを用いて，ニコニコ動画のタグを階層化する．

キーワード：共起，タグ，階層

Collaborative Creation of Hierarchical Tags in NicoNico

Takahashi Fumihiko[†1,a)]　Yamamoto Masahito[†2]　Furukawa Masashi[†2]

Abstract: Video sharing sites are so popular in Web services in all over world. One of such sites, NicoNico, which deals with video files, is successfully served in Japan. This site uses tags to classify contents of videos. It is known that the desired content is efficiently explored when tag's hierarchical information is given. This study proposes a new method to automatically generate hierarchical structure of tags in NicoNico.

Keywords: cooccurrence relation, tag, hierarchical structure

1. はじめに

国内で人気の動画共有サイトの１つに，ニコニコ動画がある．ニコニコ動画には 800 万以上の動画が蓄積されている．ユーザーは，この中から視聴する動画を探し出すが，目的の動画があらかじめ決まっているわけではな場合が多い．動画を探しだす方法として，ニコニコ動画のシステムではキーワード検索，タグ検索，関連動画やランキングなどの方法が提供されている．タグとは動画の内容を表すキーワードで，ユーザーが動画に対して複数のタグを付与することで分類が行われる．タグに用いる言葉はユーザーが自由に決める事ができ，またシステムに参加しているユーザ自身による分類なので，ユーザーの要求にあった分類が行えるという利点を持つ [1]．さらにニコニコ動画の場合すべてのユーザーで各動画のタグが共有される．このため不適切なタグが付けられても他のユーザーによって削除されるので，淘汰型のタグとも言われる [2]．

タグ検索は，検索ワードとなるタグが付いた動画を検索結果に表す検索方法である．検索結果に表示されるのは各動画のタイトルとサムネイルであり，このような動画の情報が一様に並べられる．しかし，検索結果から視聴する動画を選択するためには動画のタイトルを一つ一つ眺める必要があり，その中から好みの動画を見つけるのは検索結果は膨大であるため，非常に困難である．

このために，検索による検索結果をタグによって分類することが考えられる．検索結果が分類されることで，ユーザーは視聴する動画の候補を絞り込むことができる．また，タグを使った分類であるためユーザーのニーズに沿った分類がされることも期待される．検索結果の動画を分類する手段として，動画のタグによる関係からクラスタリングすることが考えられるが，検索のたびに検索結果の数万もの動画をクラスタリングする方法は現実的ではない．したがって，あらかじめ用意した分類基準に従って検索結果の分類を行う．

[†1] 現在，北海道大学 工学部 情報エレクトロニクス学科
Presently with Hokkaido Uniersity
[†2] 現在，北海道大学 情報科学研究科
Presently with Hokkaido Uniersity
[a)] takahashi222@complex.ist.hokudai.ac.jp

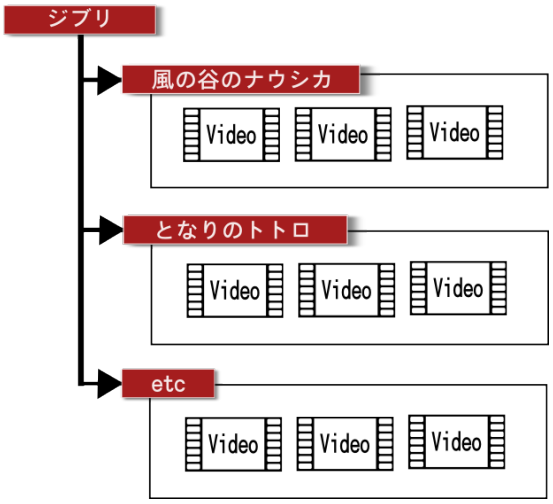

図 1 検索結果分類のイメージ

本研究では，分類基準にタグの階層構造を用いる方法を考えた．階層構造において下位概念にあるタグとは他のタグと比較して，より具体的，より局所的なタグである．この関係を用いて，検索ワードの下位概念にあたるタグそれぞれをカテゴリとして扱い，そのタグが含まれる動画をそのタグのカテゴリに分類する．この際に下位概念となるタグが含まれない動画は ”etc ” というカテゴリに分類される．この検索結果分類のイメージを図 1 に示した．矢印の向く先が下位概念となるタグであり，このイメージは ”ジブリ ” という検索ワードで検索した際の分類である．この方法の利点は，分類基準さえあらかじめ生成しておけばすでに実装されているタグ検索機能を用いることで分類が可能になるという点にある．したがって，このような分類基準となる階層構造を生成することを目的とする．

2. 関連研究

タグの関係性から階層構造を作り出すための方法が Heymann ら [3] によって提案されている．この方法はネットワーク理論的なアプローチに基づくもので，タグの共起頻度から算出されるタグ間の類似度をリンクの重みとした重み付き無向ネットワークを使って，このネットワークの中心性が高いものを上位概念になりやすいタグとして抽出する．これは，タグの関係性ネットワークにおいて重要な役割を果たすタグほど上位概念になりやすいと考えに基づいている．Heymann ら [3] が提案した階層化アルゴリズム以下のようである．

(1) すべてのタグペア $t_i, t_j \in T$ 間において類似度 S_{ij} を求める．
(2) S_{ij} を t_i，t_j 間のリンクの重みとして，タグ関係性ネットワークを生成する．
(3) 全ての t_i に対して中心性 C_i を求める．
(4) C_i の高い t_i から順にキュー構造のリストに追加する．
(5) リストの先頭から t_i を取り出し，階層構造となる木構造に追加する．
- t_i は，木構造に属するタグの中で最も S_{ij} が高い t_j の子要素として追加する．
- このとき S_{ij} が閾値 α に満たない場合は，木の根 $root$ の子要素として追加する．

(6) リストの要素がなくなるまで，5 を繰り返す．

このアルゴリズムは，タグの関係性ネットワークからエッジを除去しシンプルな木構造に置き換えることで階層化を実現するものとも言える．

村上ら [4] の研究では，細やかな動画検索を実現することを目的として，タグの共起頻度解析および ISR(Inter-section ratio) 手法によるタグの階層化を行なっている．また，この階層構造におけるタグの上下関係を表示するという方法で，ユーザーに興味の有りそうなタグを発見するためのツールを構築している．ISR 手法は，タグ t_i, t_j があるとき，

$$D(t_i) > D(t_j) \tag{1}$$

かつ，

$$\frac{|D(t_i) \cap D(t_j)|}{|D(t_j)|} > \alpha \tag{2}$$

のとき，t_i は t_j の上位タグであると定義する方法である．ここで $D(t_i)$ はタグ t_i の出現回数である．ISR 手法による階層化は，タグペアの上下を決めるだけで階層構造全体の構造は考慮していない．村上らは研究の考察で，ISR 手法による階層化はニコニコ動画のタグに適応した場合，同義語の共起が生じにくいため同義語の抽出が行えないという問題点を指摘している．また，このために他の階層化方法を適用する必要があるとも述べている．

3. ニコニコ動画

3.1 ニコニコ動画とは

ニコニコ動画とは，ドワンゴが提供する動画共有サービスである．2007 年 1 月に β 版としてサービスが開始された．再生中の動画上にコメントを付加し，共有することができるコメント機能が特徴的である．

3.2 タグ

3.2.1 フォークソノミー

ニコニコ動画ではタグによる分類が行われているが，このようなタグによる分類をフォークソノミーという．フォークソノミーは folk(民族・人々) と taxonomy(分類) を掛け合わせた造語で，リソースに対して複数のユーザーがタグを付与することで分類を行う．タグとはリソースの内容を表現するキーワードである．ユーザーはタグに自由な言葉を用いることができ，編集や消去も自由なので，多くのユーザーが分類に協力できる．したがって，従来の統制語彙による分類（タクソノミー）に比べてフォークソノミー

は分類のコストを大幅に減少させることができ，大規模なリソースに対しても有効であると考えられている．

専門家が規定した分類を使うのではなく，システムに参加しているユーザー自身が自由に言葉を選んでタグを付けられるため，ユーザーの実際の要求と言葉に適合させることができるという利点を持つ．また，多数のユーザーの主観による分類が共有されるため，他のユーザーによる分類から思いがけないリソースの発見も期待される．

一方で，フォークソノミーではタグに用いる言葉が自由に設定できるために生じるいくつかの欠点を抱えている．表記のゆらぎは，同じ属性を表すタグが複数存在するという同義語の問題である．例えば，”携帯電話 ”というタグもあれば ”携帯 ”といったタグも存在するように，どちらも同様の属性を表すにもかかわらず異なるタグとして存在する．また，１つのタグが複数の属性を表すという多義語の問題も抱える．例えば，”apple”というタグがあるときこれは「りんご」という意味と「Macintosh を販売している会社の Apple」という意味の複数の属性を持つタグが存在する．また，メタ・ノイズといった分類に意味を成さないタグが混在するために，相対的に分類が煩雑になってしまうという問題も挙げられる．

3.2.2 ニコニコ動画のタグ

ニコニコ動画のタグは，一般的なフォークソノミーと異なるいくつかの特徴がある．

タグの共有

タグが全てのユーザーで共有される．したがって，あるユーザーが追加したタグは他のユーザーによって編集・削除される可能性がある．

タグの数制限

1 つの動画に付与できるタグは最大 10 個までと決められている．このため，ユーザーが 10 個のタグが付いている動画にタグを付けたいときは他のタグを削除する必要がある．

カテゴリを表すタグ

動画を投稿する際に投稿者が選んだカテゴリに対応するタグ（カテゴリタグ）が付与される．2013 年 2 月 7 日現在は，カテゴリは 30 個ある．

タグのロック

動画投稿者は動画のタグを 5 個まで他のユーザに削除・編集できないようロック（タグロック）することができる．

このような特徴から，ニコニコ動画の場合ユーザー個人個人の分類というよりも淘汰型の分類がなされると言われている [2]．

4. タグの階層化

4.1 実験で用いるデータセット

実験で用いるデータを集めるために，ニコニコ動画によって提供されている API を用いた．2013 年 1 月 6 日頃，キーワード検索の検索結果に含まれるタグを集め，それを再びキーワード検索にかけることで再帰的にタグを集めた．また集めたタグをタグ検索に再度かけ，タグの出現回数も取得した．約 40 万件の動画から 426,937 個のタグを集めた．実験で用いるデータセットは，集めたタグを出現回数で降順にソートしその上位 100 個のタグとした．この際，”投稿者コメント ”というタグはシステムが自動で付加するタグでありユーザーによる分類ではないのでこれを除いた．実験で用いるデータセットを出現頻度で降順にソートした際の，上位 10 件と下位 10 件の一覧を表 3 に示す．

rank	name	frequency
1	ゲーム	3731236
2	実況プレイ動画	1176511
3	音楽	885621
4	歌ってみた	579147
5	エンターテイメント	360841
6	アニメ	350557
7	VOCALOID	239125
8	東方	236818
9	もっと評価されるべき	231453
10	アイドルマスター	210816
91	ニコニコインディーズ	25346
92	作ってみた	25283
93	ファイアーエムブレム	25066
94	巡音ルカ	24881
95	CM	24714
96	ニコニコ動画講座	24691
97	バイオハザード	24583
98	AxisPowersヘタリア	24368
99	ガンダム	24232
100	EXVS	24126

表 1 実験で用いるタグの出現頻度上位 10 件と下位 10 件

4.2 タグの類似度

あるタグペアが同じリソースに付与されているとき，そのタグペアは共起しているという．ニコニコ動画の場合，リソースは動画なので同じ動画に付与されているタグを共起しているという．本研究では，共起しやすいタグは階層構造の同階層の関係というよりも上位・下位の関係に近いと考えた．

しかし共起回数を類似性尺度に用いることを考えたとき，単純に出現回数の多いタグは共起回数も多くなる傾向にあるので，例え関係の薄いタグペアであっても高い類似度が出る事がある．したがって，共起回数を正規化し類似度として扱うためにダイス係数（Dice coefficient）を用い

る．ダイス係数は，共起回数に対して各集合の要素数の和で割ることでこれを正規化している．出現回数の割に共起回数が高いタグペアの類似度が高くなる傾向にある．タグ t_a が付けられた動画の集合を A，タグ t_b が付けられた動画の集合を B とすると，ダイス係数は，式 3 で表される．

$$S_d(t_a, t_b) \models \frac{|A \cap B|}{|A| + |B|} \tag{3}$$

4.2.1 類似度の比較実験

ダイス係数と他の類似性尺度でタグペアの類似度を算出し，各類似性尺度で降順にソートしたときのリストを比較する．実験で用いたタグは，前述したデータセットを用いる．比較に用いる他の類似性尺度は，共起回数・コサイン尺度を用いる．

コサイン尺度（Cosine similality）は，テキストマイニングの分野で文章間の類似度を測る際にによく利用される類似性尺度である．この尺度では，各タグの特徴ベクトルを比較して類似度を計算する．タグ t_a，t_b の特徴ベクトルをそれぞれ $\boldsymbol{v_a}$，$\boldsymbol{v_b}$ とした時，コサイン尺度は式 4 で表される．Heymann ら [3] の実験でも類似度の尺度として，コサイン尺度が採用されている．

$$S_c(t_a, t_b) \models \frac{\boldsymbol{v_a} \cdot \boldsymbol{v_b}}{||\boldsymbol{v_a}||||\boldsymbol{v_b}||} \tag{4}$$

特徴ベクトルの設定には様々な方法がある．Heymann ら [3] の実験ではリソースをベクトルとして用いている．ベクトルの長さをリソースの数として，それぞれのリソースに対してそのタグが付けられた回数をベクトルの要素としている．しかしながらニコニコ動画の場合，各リソース（動画）のタグはすべてのユーザーで共有されるため，リソースに対するタグが付与された回数が存在しない．したがって，ニコニコ動画におけるタグの特徴ベクトルの場合，共起するタグをベクトルとすることが考えられる．タグ t があるタグ t' と共起した回数をベクトルの要素 $v_{tt'}$ とし，タグの総数次元を持つベクトル $\boldsymbol{v_t}$ を定義する．この時 $v_{tt} = 0$ とする．これは，「同じタグと共起しやすいタグは類似している」という考えに基づいている．またこの共起するタグを用いたコサイン尺度は，既存研究によって同義語を見つけるための尺度として有効であることが検証されている（[5]）．

共起回数で降順に並べたリストの上位１０件を表 2 に，コサイン尺度で降順に並べたリストの上位１０件を表 3 に，ダイス係数で降順に並べたリストの上位１０件を表 4 に示す．共起回数では，関係性の高いタグペアが高い類似度を示す傾向が見られる．また，出現頻度の多いタグが高い類似度を示している．例えば，”ゲーム”タグの出現回数（付与されている動画数）はニコニコ動画全体の動画の約４０%近くに及ぶが，”ゲーム”タグを含むタグペアが上位に頻出している．コサイン尺度では，使われ方や意味の近いタグが高い類似度を示している．また，ゲームに関係するタグが上位を占める．これは”ゲーム”タグの出現回数が突出しているため，各タグの”ゲーム”タグとの共起回数も多くなり，特徴ベクトルの”ゲーム”タグ次元が強く影響していると考えられる．この結果から，出現回数の多いタグの次元に強く影響されるため，何らかの正規化が必要であるといえる．階層構造の同階層のタグを発見する際に利用できると考えられる．ダイス係数では，[”動物”・”猫”]や [”歌ってみた”・”ボカロオリジナルを歌ってみた”] のように上下関係の強いタグが高い類似度を示している．また，出現回数の依存も見られない．

tag1	tag2	cofreq
ゲーム	実況プレイ動画	1149587
ゲーム	実況プレイ	167703
アイドルマスター	IDOLM@STER	165571
ゲーム	プレイ動画	138775
歌ってみた	ボカロオリジナルを歌ってみた	136319
ゲーム	ゆっくり実況プレイ	113013
VOCALOID	初音ミク	93418
ゲーム	Xbox360	89314
ゲーム	PS3	88240
音楽	ニコニコムービーメーカー	83480

表 2 共起回数を類似度にした時の類似度の高いタグペア

tag1	tag2	cos
EXVS	戦場の絆	0.999015
バイオハザード	ゼルダの伝説	0.998876
MUGEN	戦国大戦	0.998582
戦国大戦	戦場の絆	0.998493
ゲーム実況	実況	0.998042
Wii	ファイアーエムブレム	0.997723
EXVS	戦国大戦	0.997621
実況プレイ	実況	0.99749
実況プレイ動画	戦国大戦	0.997441
格闘ゲーム	戦国大戦	0.997416

表 3 コサイン尺度を類似度にした時の類似度の高いタグペア

tag1	tag2	dice
動物	猫	0.739904
歌ってみた	ボカロオリジナルを歌ってみた	0.592351
アイドルマスター	IDOLM@STER	0.502761
ゲーム	実況プレイ動画	0.494401
三国志大戦	三国志大戦3	0.426444
R-18	エロゲ	0.363909
車載動画	バイク	0.327173
ボーダーブレイク	BORDERBREAK	0.31267
演奏してみた	弾いてみた	0.283685
VOCALOID	初音ミク	0.282569

表 4 ダイス係数を類似度にした時の類似度の高いタグペア

この結果から，本研究で上下関係の強いタグペアに高い値を出す指標としてダイス係数を

4.3 上位概念になりやすいタグ

階層構造の上位層になりやすいタグは，他のタグに比べてより包括的，より抽象的な意味を持つタグである．Heymann ら [3] の研究では，上位概念になりやすい尺度として関係性ネットワークの中心性尺度を用いている．このネットワークは，上位下位概念になりやすいタグの尺度と

tag	degree		close		rank
		rank		rank	difference
ボカロオリジナルを歌ってみた	0.6326937	92	0.0088087	70	22
VOCALOID	0.9663017	17	0.0101677	5	12
エンターテイメント	0.9665521	16	0.0101651	6	10
実況プレイ動画	0.9407374	28	0.0099905	19	9
動物	0.9203154	38	0.0097856	29	9
猫	0.8916864	44	0.0096868	36	8
エロゲ	0.8345339	58	0.0091813	50	8
車載動画	0.7994531	69	0.0089501	61	8
ゲーム	0.972854	8	0.010279	1	7
IDOLM@STER	0.8789117	45	0.0096762	38	7
ニコニコ動画講座	0.9666278	15	0.0099853	20	-5
神曲	0.9465807	26	0.0097569	31	-5
ガンダム	0.926923	34	0.0096415	40	-6
バイオハザード	0.868414	49	0.0091428	55	-6
例のアレ	0.9855125	1	0.010147	8	-7
MUGEN	0.8288318	59	0.0088412	66	-7
テイルズ	0.8181947	64	0.0087852	72	-8
UTAU	0.8161638	66	0.0087685	74	-8
高画質	0.9756524	6	0.0100461	16	-10
ラジオ	0.9774236	5	0.0100275	17	-12

表 5 次数中心性と近接中心性の比較

して用いた類似度をリンクの重みとしたネットワークである．関係性ネットワークにおいて重要な役割を果たすタグほど上位概念になりやすいと考え，重要な度合いの尺度として中心性指標を用いた．中心性指標にはいくつか種類があるが，Heymann ら [3] は実験で近接中心性を用いている．

近接中心性（Closeness centrality）は，「他のノードと距離が近いほど中心性が高い」という考えに基づく指標である．ネットワーク理論における距離とは，最短経路長を指す．ノード v_i からあるノード v_j への最短経路長を $d_G(v_i, v_j)$ としたとき，ノード v_i から $|V|-1$ 個の他のノードへの最短経路長の長さの平均 L_i は，

$$L_i = \frac{\sum_{j=1, i\neq j}^{|V|} d_G(v_i, v_j)}{|V|-1} \tag{5}$$

であり，近接中心性はこの逆数で表される (式 6)．

$$C_C(v_i) = \frac{1}{L_i} \tag{6}$$

近接中心性の高いのノードは他のどのノードからも少ないホップ数でたどり着けるという意味で中心的な役割をしているといえる．つまり，近接中心性の高いタグは他の様々なタグと関係が強いと考えられる．多くのタグに関係するタグは，より包括的，より抽象的な意味を持つタグであると考えられる．また，局所的な関係だけではなく全てのタグペアの関係を考慮している．

4.3.1 上位概念へのなりやすさの尺度の比較実験

多くのタグに関係するタグがより上位概念になりやすいと考えるならば，そのタグに対する他のタグとの類似度の合計が高いタグが上位概念になりやすいとも考えられる．この尺度は関係性ネットワークにおける次数中心性に一致する．したがって，次数中心性の高いタグと近接中心性の高いタグを比較する．関係性ネットワークを，前述したダイス係数を用いてリンクの重み（非類似度）を $\{1 - S_d(t_a, t_b)\}$ として生成したネットワークを用いる．関係性ネットワークで次数中心性と近接中心性を計算し，次数中心性と近接中心性それぞれでタグを降順にソートしたリストを作成し比較する．またこのリストを比較するために，タグごとに次数中心性における順位から近接中心性における順位を引いて，その値でソートしたリストを作成する．リストの上位１０件と下位１０件を表 5 に示した．出現回数を用いた場合に比べて，近接中心性を用いた場合の方がより上位に位置するタグが，このリストの上位に来る．

カテゴリタグはシステムが提示するカテゴリに分類するためのタグであるため，上位概念になりやすいと考えられる．すると近接中心性では，”エンターテイメント”や”ゲーム”，”VOCALOID”といったカテゴリタグが次数中心性よりも上位に位置している．

この結果から，本研究では上位概念へのなりやすさの指標として近接中心性を用いることにする．

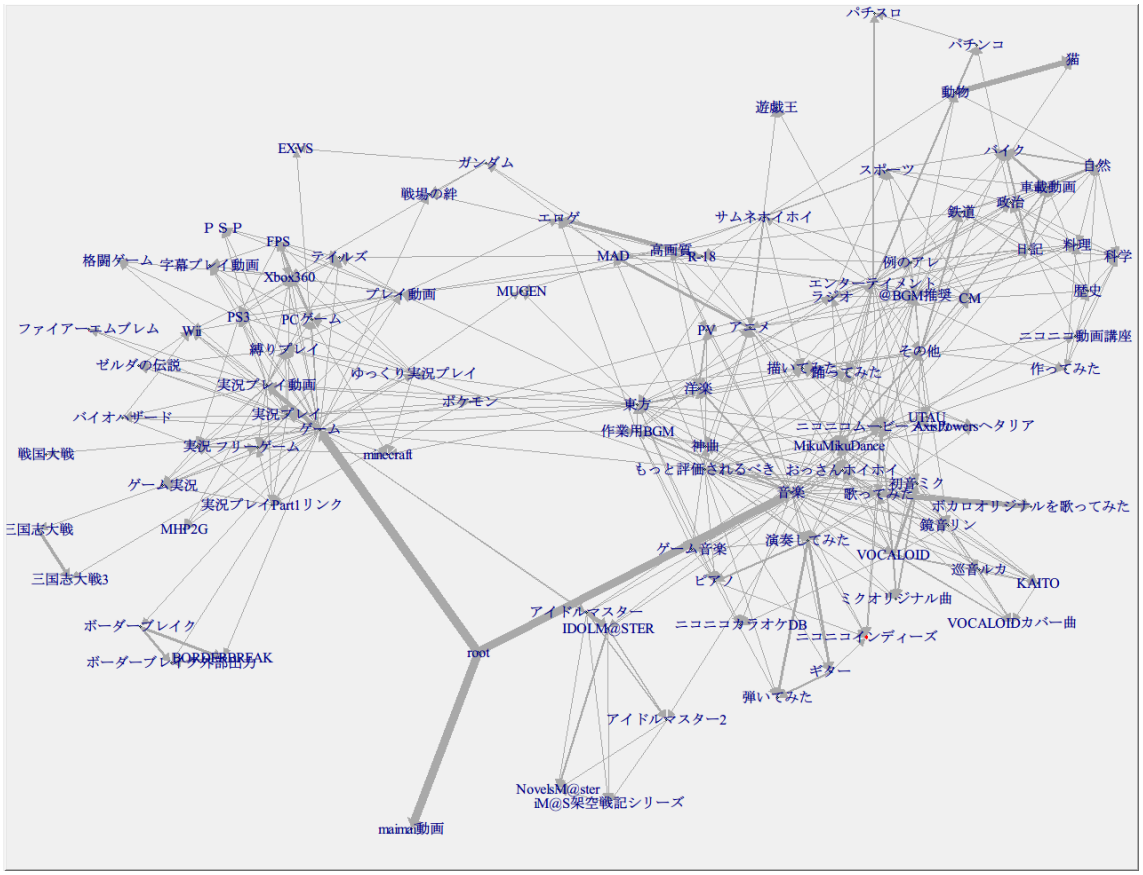

図 2 出現頻度と近接中心性の比較

4.4 タグの階層構造化

Heymann ら [3] の研究で提案されている階層化方法に沿ってタグを階層化を行う．しかしながら，この階層化方法は階層を木構造で作るという制約を持つ．木構造でタグを配置するとタグの上位タグは必ず 1 つに限定され，このようにして決定された木構造には不自然な箇所が発生すると既存研究 [6] で指摘されている．本研究でも検索結果の分類を行う際には，分類のカテゴリとなる下位タグは関係が強ければ多く存在しても構わない．したがって，階層構造にタグを追加する際に閾値以上の類似度を持つ上位タグ全てにタグを追加する．

生成した階層構造を，2 次元空間上に有向ネットワークとして配置した結果を図 2 に示す．リンクの矢印は上位タグから下位タグに向かっている．

5. おわりに

本研究では，動画の検索結果の分類で用いる分類基準とするために，Heymann ら [3] によって提案されたタグの階層化方法を拡張して，複数の上位タグや複数の下位タグをそれぞれのタグが持つようなタグの階層構造を生成した．

今後の課題は，この階層構造に沿った分類を行った場合の結果を評価することや動画に対する意見や感想のような分類に寄与しないタグの処理などが挙げられる．

参考文献

[1] 緑川信之: フォークソノミーの新奇性はどこにあるのか, 情報の科学と技術 57(5), 238-243, 2007-05-01 (2007).
[2] 伊藤聖修，鈴木育男，山本雅人，古川正志: ニコニコ動画におけるタグ共起ネットワークの特徴抽出, 第８０回人工知能学会知識ベースシステム研究会 (2008).
[3] Paul Heymann, Hector Garcia-Molina：*Collaborative creation of communalhierarchical taxsonomies in social tagging systems*, Techniqual Report 2006-10, Stanford University, April 2006.
[4] 村上直至, 伊東栄典: 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).
[5] Ciro Cattuto, Dominik Benz, Andreas Hotho, Gerd Stumme: *Semantic analysis of tag similality measures in collaborative tagging systems*, (2008).
[6] Christopher H. Brooks and Nancy Montanez: *Improved annotation of the blogosphere via autotagging and hierarchical clustering*, In Proc. WWW, pp. 625—632, New York, NY, USA. ACM Press (2006).