

動画投稿サイトで付与された 動画タグの階層化

村上直至[†] 伊東栄典[‡]

あらまし 国内で人気の高い動画投稿閲覧サービスに YouTube とニコニコ動画がある。これらのサービスでは膨大な動画が蓄積されており、利用者が好みの動画を探すのは容易ではない。サービス運営側は検索語と並び替えによる動画検索を提示しているものの、細やかな動画検索は実現できていない。そこで、多数の視聴者が動画に付与するタグを用いた動画分類を考える。本研究ではニコニコ動画を対象とし、動画分類や類似動画提示のために、タグの出現頻度および共起出現を用いたタグの階層化を行った。階層化には ISR(Inter section ratio)手法を用いた。

キーワード 動画投稿サイト、動画分類、タグ、folksonomy、階層化

Tag hierarchy analysis in Nicovideo movie service

Naomichi Murakami[†] and Eisuke Ito[‡]

Abstract It is very difficult to find favorite movie from recent popular movie uploading and streaming service such as YouTube and Nicovideo, because there are a lot of movies in those sites. Movies are given not only metadata such as title, creator, upload date and movie length by the creator, but also annotations such as tags and comments by viewers. Those annotations may be used as folksonomy based data mining, because viewer may present his/her sympathy in annotation. In this paper, we analyzed hierarchy of tags using frequency of tag.

Keyword Movie service, Nicovideo, tag, categorization, folksonomy, hierarchy

1. はじめに

近年人気のある YouTube[1]やニコニコ動画[2]などの利用者投稿型動画共有サービスでは、毎日多数の動画が投稿されており、サイト全体には膨大な数の動画が蓄積されている。膨大な動画から、利用者が求める面白い動画を探すためには検索システムが必要である。

我々は、ニコニコ動画を対象に動画検索システムについて研究している[4,5]。ニコニコ動画は2006年12月12日に運営を開始した、視聴者投稿型の動画配信サービスである。運営開始から4年半以上経過した2010年9月末現在、420万件を超える動画が投稿されている。会員数も膨大であり、ID登録者数は1800万を超えており、そのうち有料のプレミアム会員も100万を超えている[3]。

現在、ニコニコ動画では(YouTubeでも)、検索語入力および検索語に対する適合動画の並び替え手法の選択による動画検索が提供されている。並び替えには、投稿日時、再生回数、評価値によるものが提供されている。他に、カテゴリ指定による検索対象動画の制限機能もある。

これらの手法は、検索対象となる動画が明確な場合は有効である。しかし、なんとなく面白い動画を探したいような対象を特定していない場合、検索語を上手く選定できないため検索がうまくいかない。なんとなく面白い動画を探したい場合、動画の自動分類や、検索語に近い動画の提示、閲覧者の好みや閲覧履歴等のプロフィールに適した動画の提示といった機能が必要である。

現在のニコニコ動画でも(YouTubeでも)、ある動画に対する類似動画・関連動画の提示機能が提供されている。関連動画の提示方法は明示されていないものの、視聴者の共起アクセスに基づく手法を用いているものと思われる。しかし、視聴した動画の関連動画から自分の好みの動画を見つけることは困難である。関連動画へのリンクを辿りながら好みの動画を見つける確率は、動画空間の膨大さから考えて低い。

動画分類や類似動画検索を行う場合、利用者が付与するタグを Folksonomy 的なマイニング資源として用いれば、タグによる動画特徴付けが可能である。動画投稿者および多数の視聴者から付与されるタグは、動画の分野や面白さを表現する情報となっている。タグの解析で動画投稿者および視聴者が、その動画をどの分野として指定したいかを汲み取ることができれば、動画の検索や、類似動画の推薦および分類に有効である。

ニコニコ動画では、動画に付与されたタグについて、タグクラウドを提示している。タグクラウドは、タグ数が少ない場合か、人気の高いメジャーなものを見つけるには有用である。しかし、膨大な動画から自分の好みに合う動画に付与さ

[†]九州大学大学院システム情報科学研究院, 福岡市
Graduate School of Information Science and Electrical Engineering, Kyushu University,
Fukuoka-shi, 812-8581 Japan

[‡]九州大学情報基盤研究開発センター, 福岡市
Research Institute for Information Technology, Kyushu University, Fukuoka-shi, 812-8581
Japan

E-mail: [†] 2IE10084E@s.kyushu-u.ac.jp, [‡] ito.eisuke.523@m.kyushu-u.ac.jp

れたタグを見つけることは、タグクラウドだけでは困難である。何らかの形でタグを構造化し、構造に基づいてのタグ検索および動画検索を実現することが必要である。

本稿では、動画分類や類似動画提示のために、タグの出現頻度および共起出現を用いたタグの階層化を行った。階層化には ISR(Inter-section ratio)手法を用いた。本論文の構成を述べる。2節では関連研究について述べる。3節では「音楽」タグの付いた動画集合における、タグの出現頻度解析について述べる。4節では、タグの階層化手法、実データに適用した試作システム、および結果の考察を行う。5節でまとめと今後の課題を述べる。

2. 関連研究

タグの階層化については、ソーシャルブックマーク(social bookmark)を対象にした研究が広く行われてきた。R. Li らはソーシャルブックマークのタグについて、ISR 手法を用いた階層化を行い、その結果を用いたタグ閲覧システム ELSABer を提案している[6]。また、手法をソーシャルブックマークサービス Flickr におけるタグに適用した試作システムを提示している。百田も同じく、ソーシャルブックマークサービスである「はてなブックマーク」に対し、タグの階層構造を視覚化した[7]。下司らは共起頻度による概念グラフを構築した後、隣接する隣接上位ノードを用いた推薦手法に応用し[8]、特許[9]を取得している。一般に Web 上のデータは単語の出現頻度に大きなばらつきがある。榊らは出現頻度のばらつきを正規化するため、本稿で用いた Inter-section ratio の代わりに χ^2 値を用いて親子関係を導出する手法を提案した[10]。

2009 年 1 月には、自発的な研究を行っていた研究者が、ニコニコ動画データ分析研究発表会[11]を開催した。末永匡は、榊剛史らの論文[7]に倣い、 χ^2 値を用いた単語間の関連度の指標を使って、ニコニコ動画の Folksonomy 的なタグ間の関連度を求めている[12]。

動画の分類にはタグ以外にも、投稿者間のつながりを用いる方法がある。濱崎らはニコニコ動画における初音ミク動画コミュニティを対象として動画作者ネットワークを分析した[13]。

Ido Guy らはアイテムベースの推薦において、利用者間の距離を用いた場合とタグ間の距離を用いた場合についての比較研究を行った。論文[14]ではタグ間の距離を用いる推薦の方が良い性能を出し、且つ二者を組み合わせることで更なる性能の向上が見られたことを報告している。本稿ではタグについてのみ着目したが、利用者の作るプレイリストを複合して考えるとより興味深い階層化が可能かもしれない。

3. タグの頻度分析

3.1 ニコニコ動画のタグ

ニコニコ動画へ投稿された動画にはメタデータとして「投稿者アカウント、動画 ID、タイトル、説明文、投稿日時、タグ、サムネイル画像、動画の長さ、コメント、再生回数、マイリスト登録数」が付与される。このうち、タイトル、説明文、投稿日時、タグの一部、コメントの一部、サムネイル画像は投稿者が指定する。動画長は投稿動画から自動的に決まる。図 1 に付与される情報の概要を示す。

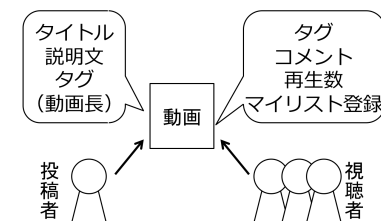


図 1 動画に付与されるデータ

次にタグについて述べる。図 3.2 に示すように、タグは画面では動画のすぐ上部に提示される。最高 10 個のタグが動画に付けられる。動画投稿者は、投稿した動画にタグを付ける事と、タグをロックする(他者が編集出来ない状態にする)ことが出来る。ロックされていないタグは、視聴者により自由に上書きできる。自由な上書きは可能であるが、ある程度の期間が経過して評価の定まった動画では、タグが固定化する傾向がある。

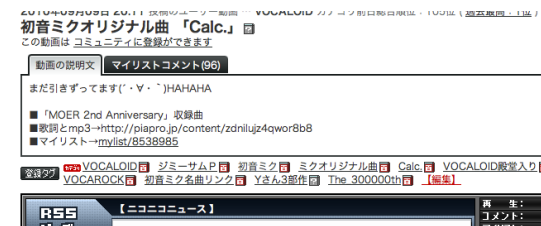


図 2 タグの提示

図 2 のタグの横にある「百」アイコンは、ニコニコ大百科に説明エントリが存在する場合に表示されるリンクアイコンである。ニコニコ大百科は、Wikipedia

的な単語説明サイトで、利用者の自由編集による説明文が記載されている。

タグの一部にカテゴリタグがある。カテゴリタグとは、ニコニコ動画の運営側が指定するタグである。ニコニコ動画では、代表的なタグが、動画の大分類となるカテゴリタグに指定されている。動画投稿者は、自分の動画を分類されやすくするため、カテゴリタグをロック状態で付与することが多い。2010年9月末のカテゴリタグは以下の通りである。「エンターテインメント、音楽、スポーツ、動物、料理、日記、自然、科学、歴史、ラジオ、ニコニコ動画講座政治、歌ってみた、演奏してみた、踊ってみた、描いてみた、ニコニコ技術部、アニメ、ゲーム、アイドルマスター、東方、VOCALOID、例のアレ、その他、R-18」

3.2 音楽カテゴリ動画のタグ頻度解析

「音楽」タグがついた動画（約37万件）を対象として、タグの頻度解析を行った。ニコニコ動画には2010年9月末で420万以上の動画が蓄積されている。その全てを分析することは困難である。そこで、最初のカテゴリタグである「音楽」タグが付与された動画を対象とした。

2009年10月28日より11月5日までの間、「音楽」タグが付与された動画373,468件のメタデータを取得した。最初に「音楽」タグを持つ動画の動画IDを取得し、各動画に付随するメタデータ収集した。メタデータ収集にはニコニコ動画から公式に提供されているAPI[3]を利用した。動画IDリスト収集とメタデータ収集に時間差があるため、アノテーション取得の際に「音楽」タグが消えてしまっていた動画もあった。

ID取得時には「音楽」タグを持つ動画が373,468個であったが、メタデータを取得した結果では6,690個の動画から「音楽」タグが消えていた。残りの317,222個では「音楽」タグがロックされており、49,457個では非ロック状態で「音楽」タグが付けられていた。

3.3 全タグ

収集した動画集合には、一意なタグが355,872個存在した。このうち70.6%にあたる251,229個のタグは「音楽」タグと一度しか共起しなかった。一度だけ共起したタグは、他の視聴者をうならせるネタ的意味合いの強いタグや、付与された動画固有の説明文となっているタグが散見された。これらのタグは検索における効果は薄いものの、動画コンテンツを盛り上げる要素となっている。

表1に上位10位のタグと頻度を示す。1位の「ニコニコムービーメーカー」と第5位の「投稿者コメント」は、ニコニコ動画のシステムが自動付与するタグである。

1位の「ニコニコムービーメーカー」は、静止画と音声を組合せて動画を作る無償ツールである。このツール経由の投稿動画には、このタグがロック状態で自動付与される。「ニコニコムービーメーカー」が高頻度で出現した理由は、対象

が「音楽」タグを持つ動画であるからであろう。音楽動画では音声に比べ映像はさほど重視されないため、作曲者が気軽に投稿できるツールの利用頻度が高いと思われる。

5位の「投稿者コメント」タグは、動画投稿者がロックしたコメントを入れた場合に自動付与される。なお、自動付与タグは投稿者が自発的にタグロックを解除することで消去できる。

自動付与タグでない高出現頻度タグには「VOCALOID」「初音ミク」「ミクオリジナル曲」等のボーカロイド関連タグがある。7位の「東方」はゲームソフトのシリーズ名で、特にBGMの評判が高い。これらは2009年後半までの人気分野を示している。

表1 出現頻度上位10位のタグ

順位	頻度	タグ
1	49007	ニコニコムービーメーカー
2	34236	VOCALOID
3	31002	作業用BGM
4	30255	初音ミク
5	29689	投稿者コメント
6	27950	洋楽
7	17100	東方
8	12993	ミクオリジナル曲
9	11943	もっと評価されるべき
10	9826	邦楽

3.4 ロックタグ

動画投稿者によりロックされたタグの頻度を解析した。一意なロックタグは66,180個で、そのうち40,304個（60.9%）は頻度1であった。動画投稿者がロックするタグでは、検索に有用なタグに加え、先に述べたネタ的意味合いの強いものもある。

表2 出現頻度上位10位のロックタグ

順位	頻度	タグ
1	44221	ニコニコムービーメーカー
2	27059	投稿者コメント
3	12872	初音ミク
4	9422	作業用BGM

5	7123	VOCALOID
6	6989	東方
7	6895	洋楽
8	6762	@BGM 推奨
9	3766	ミクオリジナル曲
10	3086	KAITO

3.5 非ロックタグ

非ロックタグの頻度を分析した。一意なロックタグは 327,984 個で、そのうち 237,029 個 (72.3%) は頻度 1 であった。

9 位と第 10 位を除く他のタグは、表 1 の上位 10 位に入っている。「もっと評価されるべき」と「邦楽」は、ほとんど非ロック状態で出現していたことは特長的である。「洋楽」タグのロック率が 24.7%であるのに対して、「邦楽」タグのロック率が 4.9%であったことは興味深い。

非ロックタグは視聴者によって付与される。全視聴者により編集可能であるため、削除されることもある。その中で、削除されずに残ったタグは、その動画への視聴者の思い、共感が含まれていると考えられる。例えば、「もっと評価されるべき」タグは、良質動画にも関わらず再生数等の評価が低い、あるいは存在自体認知されていないと感じられた場合に、視聴者側から自然発生的に付与されていると考えられる。

表 3 出現頻度上位 10 位の非ロックタグ

順位	頻度	タグ
1	27113	VOCALOID
2	21600	作業用 BGM
3	21053	洋楽
4	17383	初音ミク
5	11771	もっと評価されるべき
6	10111	東方
7	9445	邦楽
8	9227	ミクオリジナル曲
9	7140	PV
10	6652	VOCALOID カバー曲

4. タグの階層化

4.1 ISR 手法

タグの階層化には様々な方法がある。今回は ISR(Inter-section ratio)手法を用いた。ISR では単語 u, v の上下関係を頻度で以下のように定義する。

u is upper term of v

$$\Leftrightarrow D(u) > D(v) \quad \text{and} \quad \frac{|D(u) \cap D(v)|}{|D(v)|} > \alpha$$

ここで $D(u)$ は単語 u の文書頻度である。単語 u は v より出現頻度 (文書頻度) が高く、かつ v が出現する文書中での、 u の共起確率 (ISR 値) が α を超える場合、 u は v の上位であるとする。

この ISR 手法をニコニコ動画のメタデータを対象に適用した。動画一件のメタデータを文書とし、メタデータ内のタグを単語として扱う。なお、一つの動画でタグは重複出現しないため、タグの単語頻度 (Term Frequency) と文書頻度 (Document Frequency) は同じ値となる。

例えば「VOCALOID」と「初音ミク」の場合、「VOCALOID」の出現頻度が高く、かつ「初音ミク」とタグ付けされた動画の 60%では、「VOCALOID」タグも付けられている。従って ISR 手法では「VOCALOID > 初音ミク」とする。

実際、「VOCALOID」はヤマハ社が開発・販売する音声合成エンジンで、「初音ミク」はクリプトン・フューチャー・メディア社が販売する音声合成ソフト (およびキャラクター名) であり、内部のエンジンとして VOCALOID を用いている。そのため概念的な上下関係にも合致する。

ここで重要になるのは閾値 α の値である。 α を高く設定した場合、適合率の高い階層構造を作ることができる。しかしながら、ISR 値の高いタグは検索者にとって既知である可能性も同様に高い。それに比べ ISR 値は低い α でない、一度でも共起したタグは検索者にとって意外性を持っている。これらのタグは適合率の高い階層構造を構築する上ではノイズにしかならないものの、検索面からの利用価値は十分にある。従って後述するシステムでは α による親子関係の切り捨てを行わなかった。代わりに親子関係にあるタグの上位 10 件ずつを表示することで、検索クエリに対して重要なタグのみを表示できるようにした。

4.2 試作システム

タグの上下関係を表示するシステム Tag More [a]を試作した。試作システムでも、先に述べた「音楽」タグがついた動画 373,265 万件のメタデータを対象としている。システム構成の概要を表 4.1 と図 4.1 に示す。また利用画面を図に示す。

a) <http://joe.cc.kyushu-u.ac.jp/~murakami/seemore/tagmoreindex.html>

表 4 システム構成

OS	MacOS X
Web	Web Apache 2.2
言語	Ruby 1.8.6
RDB	SQLite 3

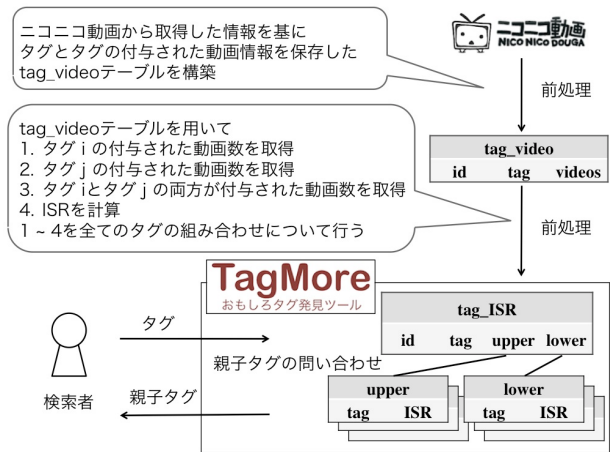


図 3 システム構成

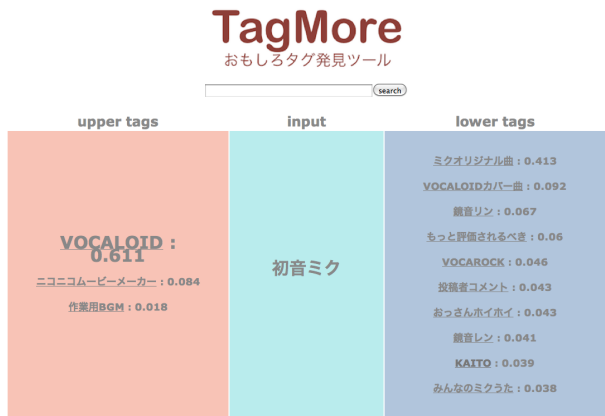


図 4 TagMore Web 利用画面

5. 考察

ISR 手法でタグを階層化した試作システムについて議論する。Adam Mathes は Folksonomy なタグ資源が、従来の分類法である Taxonomy に見られるような統制語彙を持たないことから、非階層性と言語的あいまいさを弱点に有すると指摘している[15]。ニコニコ動画の提供するタグクラウドはこれらの弱点を解決しなかったために、タグの持つ意味的な繋がりが分からなかった。本章ではこの二点から試作システムを評価する。

試作システムはタグの階層化に成功している。タグには ISR 値が 0.5 以上の関連タグがひとつ以上存在するタグと、そうでないタグがあり、それぞれにタグの特徴が出ていた。「VOCALOID」は出現頻度二位のタグで、「初音ミク」を ISR 値 0.5 で子タグに持ち、前者に属するタグである。この関連性は直感に反しない。「ニコニコムービーメーカー」は単独の出現頻度が最も高いタグで、後者に属するタグである。このタグは幅広いタグと共起する性質があるために、ISR 値が高い子タグを持たない。このことは「ニコニコムービーメーカー」タグによる検索が、動画ジャンルに対しての絞り込み効果をあまり持たないということも同時に示している。

また、今回閾値による切り捨てを行わなかった ISR 値の低いタグも良い働きをした。「初音ミク」タグは「ニコニコムービーメーカー」と同じく、様々なジャンルの音楽に用いられるツールであったため、ISR 値が 0.5 を超える子タグを持たない。しかし、ISR 値は低いながらも「ミクオリジナル曲」「VOCALOID カバー曲」タグなどを「初音ミク」から発見できた。これらのタグは、初音ミクが使用されている多数の分野の中でも特に人気のある分野を表していると考えられる。

一方で言語的あいまいさの解決には問題が残った。ニコニコ動画では一つの動画に対し 10 種類までしかタグ付けを行うことができないため、異なる二つの同義語が共起することが少ない。例えば 3 つのタグ「VOCALOID」「vocaloid」「ボーカロイド」は同義語であると思われるが、多くの場合はそれぞれが単独に出現する。表 5、表 6 にはこれら 3 つのタグに関して単独での出現頻度と共起数をそれぞれ示した。ISR 値が高くなるのは「VOCALOID」を親タグとしたときの「ボーカロイド」のみで、「VOCALOID」と「vocaloid」に至っては一度も共起しなかった。

表 5 タグの単独出現数

VOCALOID	Vocaloid	ボーカロイド
34,236	961	189

表 6 タグの共起出現数

VOCALOID	VOCALOID		
vocaloid	0	vocaloid	
ボーカロイド	115	10	ボーカロイド

参考文献

- 1) YouTube - Broadcast Yourself, <http://youtube.com> (accessed 2010.10.15)
- 2) ニコニコ動画, <http://nicovideo.jp> (accessed 2010.10.15)
- 3) ガジェット通信: ニコニコ大会議 2010 夏新サービスや新機能実装, <http://getnews.jp/archives/74530> (accessed 2010.10.15)
- 4) 村上直至, 伊東栄典: 視聴者コメントに基づく動画検索手法, 電子情報通信学会 2010 年総合大会, D-4-2 (page D-24), Mar. 2010.
- 5) 村上直至, 伊東栄典: 共感度検索のための動画アノテーション分析, FIT2010 (第 9 回情報科学技術フォーラム), F-002, 第 2 分冊, pp.343-348, Sep., 2010.
- 6) Rui Li, Shenghua Bao, Ben Fei, Zhong Su, Yong Yu: Towards Effective Browsing of Large Scale Social Annotations, WWW2007, pp.943-952, 2007.
- 7) 百田信, 伊東栄典: ソーシャルブックマークに基づく情報発見, 電子情報通信学会 DEWS2008, pp.11-15, Mar., 2008.
- 8) 下司義寛, 廣川佐千男: 概念グラフを使った推薦, 電子情報通信学会第 9 回 Web インテリジェンスとインタラクション研究会, Jul., 2007.
- 9) 廣川 佐千男, 下司 義寛, 和多 太樹: データマップ作成サーバ, データマップ作成方法, およびデータマップ作成プログラム, 特許公開 2007-257420, 2007.
- 10) 榎剛史, 松尾豊, 石塚満: Web 上の情報を用いた関連語のシソーラス構築について, 自然言語処理, Vol.14, No.2, pp.3-31, Apr., 2007.
- 11) CodeZine: 膨大なデータを分析して見えてくること, ニコニコ動画データ分析研究発表会, <http://codezine.jp/article/detail/3516>, (accessed 2010.10.15)
- 12) 末永匡: χ^2 乗値を関連度としたニコニコ動画関連タグネットワークの解析, <http://www.hotdocs.jp/file/178654>, (accessed 2010.10.15)
- 13) 濱崎雅弘, 武田英明, 西村拓一: 動画共有サイトにおける大規模な協調的創造活動の創発のネットワーク分析 -ニコニコ動画における初音ミク動画コミュニティを対象として-, 人工知能学会論文誌, Vol. 25, No. 1, pp.157-167, 2010.
- 14) I. Guy, N. Zwerdling, I. Ronen, D. Carmel, E. Uziel: Social media recommendation based on people and tags, ACM SIGIR2010, pp.194-201, 2010.
- 15) A.Mathes: Folksonomies - Cooperative Classification and Communication Through Shared Metadata, <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>, Dec., 2004

一般的なソーシャルブックマークサービスでは, タグ付けモデルに制限がなく, 同義語の ISR 値も高くなることが論文[7]で報告されている. 他方, ニコニコ動画では 10 個という制限あるタグ資源を有効に使う為に, 同義語の共起は淘汰されていると思われる. このことから, タグ付け制限を持った Folksonomy における同義語の発見には, ISR 以外の手法も必要であることが分かった.

6. おわりに

本論文ではニコニコ動画を対象に, 視聴者から付与されたタグの階層化を行った. 「音楽」タグが付与された約 37 万件の動画のメタデータを取得し, タグの頻度解析, 共起頻度, および ISR 手法によるタグの階層化を行った. 実データを用いた試作システムも構築した.

今後, タグ階層関係を用いて類似動画を提示する機能を実現したい. 今回 ISR 手法を用いてタグ間の階層関係および類似関係を計算している. このデータを用いて, タグに基づく関連動画提示を実現し, ニコニコ動画が提供する関連動画との, 品質比較を行いたい.

また, タグと視聴者コメントの出現単語との関係を調査したい. ニコニコ動画の特長は独自のコメントシステムにある. ニコニコ動画では, 視聴者がコメントを動画の再生時間に沿って入力でき, 動画再生時にコメントも提示される. 視聴者が動画へ付与するコメントには, しばしば視聴者が動画に向ける感情の切れ端が含まれ, 視聴者が楽しそうに動画を閲覧している様子がコメントから伝わる. コメントの出現単語には, 動画内容に応じた特定単語が頻出する場合もあるし, 分野 (カテゴリ) に固有の単語が出現する場合もある. 感情を表す単語, 例えば笑いを表現する 'w' が, 頻出する場合もある. これらの関係を算出し, さらに品質の高い動画検索を実現したい.