D-009

動画リストを用いた動画共有サイトにおける動画検索精度の向上

Precision Improvement of Video Search in Video Sharing Site using Public Video Lists

西 友規† 山口 実靖† 小林 亜樹†
Yuki Nishi Saneyasu Yamaguchi Aki Kobayashi

1．はじめに

インターネット上の動画共有サービスが普及し[1]，多くの動画が動画共有サイトで共有されている．しかし，動画共有サイトや Web 検索エンジンで提供されている動画の単語検索機能の精度は必ずしも十分とは言えない[2]．よって，動画共有サイトにおける単語による動画検索精度の向上は重要な課題の一つと考えることができる．

また，Web 空間からのコミュニティ抽出に関しては非常に多くの研究成果が得られており[3,4,5,6,7,8]，これらを動画共有サイト内の動画検索に応用することで，より良い検索を実現できると期待できる．

本稿では，まず動画共有サイトの機能である「タグ」と「動画リスト」について説明する．次に，既存研究である Web コミュニティの抽出手法と TF-IDF[9]，動画検索に関する研究を紹介する．そして，Web コミュニティ抽出手法と TF-IDF を用いて動画コミュニティを抽出する手法と，それを用いた動画検索手法を提案する．最後に，評価実験の結果を示し提案手法の有効性を示す．

2．動画共有サイトの機能

2.1 タグ

図 1 に動画共有サイトにおける動画と動画に付与されているタグのモデルを示す．多くの動画共有サイトでは，各動画の特徴を表す文字列をタグとして動画に対して付与することができる．例えばチャーハンの調理の動画なら，「チャーハン」や「料理」などのタグが付与されると予想される．多くの場合，タグは動画の特徴を表しており，動画の検索，分類，説明などに利用されている．タグを用いることにより，指定のタグが登録されている動画のみを検索したり，注目している動画と関連性のある動画を検索したりすることが可能となる．

2.2 動画リスト

多くの動画共有サイトで，ユーザが指定した動画群を“動画リスト”として公開する機能が提供されている．動画リストは各ユーザが自由に作成することができるが，リスト内の動画同士には関連があることが多いと期待することができる．

図 2 に動画リストのモデルを示す．各動画リストには 1 個以上の動画が登録されており，各動画にはその動画の特徴を表すタグが付与されている．

3．既存研究

3.1 Web コミュニティ抽出

Web 空間の中には共通の話題を有する Web ページ群が

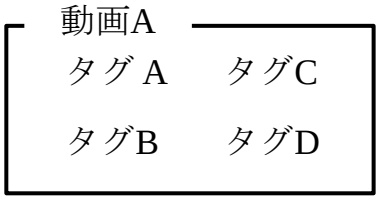

図 1 動画とタグ

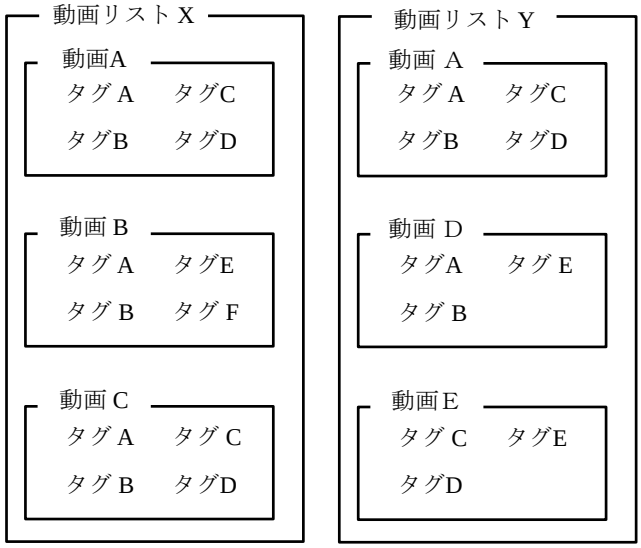

図 2 動画リスト

存在し，共通の話題を有するページ群を“Web コミュニティ”と呼ぶことができる．以下に，Web 空間からの Web コミュニティ抽出に関する研究を示す．

Kumar らは，Web ページ間のリンクグラフにおいて共通参照により構成される 2 部グラフが存在し，それらは共通の話題を持つとした．そして，密な 2 部グラフを Web コミュニティと定義した[3]．また，Web ページ群から完全 2 部グラフを抽出する手法を提案している．

Reddy らは，Web コミュニティを密な 2 部グラフ(DBG: Dense Bipartite Graph)とし，完全 2 部グラフでなく指定数以上のリンクを有する緩い共引用(Relax_cocite)による 2 部グラフの発見による Web コミュニティの抽出手法を提案している[4].

村田は，Web の検索エンジンを用いてリンクの共起の完全 2 部グラフを抽出する手法を提案している[5].

斉田らは，Web コミュニティ内のメンバが持つ話題が，シードの Web ページ群のトピックスから離れている量を表す距離量の概念を加えて精度と網羅性を向上させる PlugDBS[6]を提案している．

これらの手法では，図 3 の様に特定の話題を持つ Web ペ

†工学院大学大学院　工学研究科　電気・電子工学専攻
Electrical Engineering and Electronics, Kogakuin University Graduate School

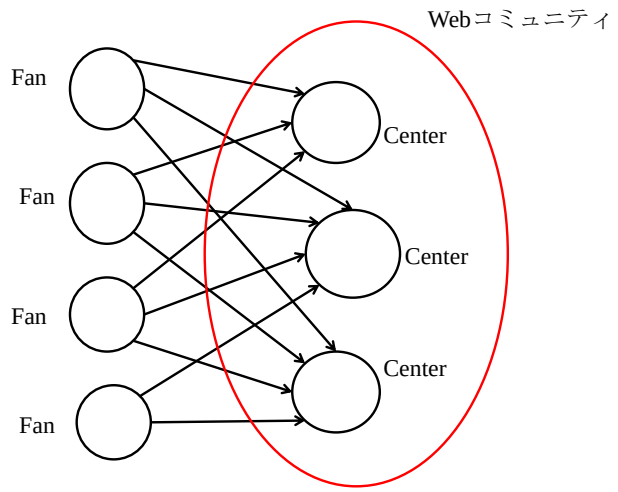

図 3 Center と Fan の関係

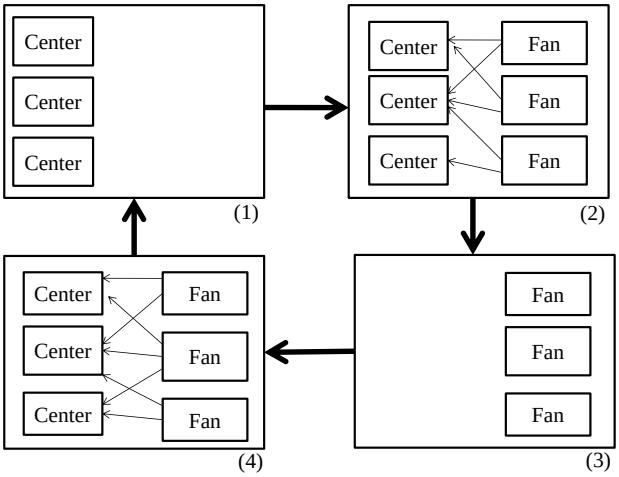

図 4 Web コミュニィテ抽出手法

ージを Center とし，Center の集合を Web コミュニティと考える．また，Center の Web ページ群に対して多数のリンクを出している Web ページを Fan としている．この Center の集合と Fan の集合を用いて二部グラフを作成し，Center 集合からの Fan 集合の作成および Fan 集合からの Center 集合の作成を繰り返し Center 集合の改善を行っていく．

図 4 にこれらの手法における Web コミュニティ抽出の手順の概要を示す．最初に初期 Center 集合あるいは初期の Fan 集合を与える．Center 集合を与えた場合，その Center 集合の内のすべて(あるいは指定数以上)のページに対してリンクを出しているページを Fan として抽出する．次に，この Fan 集合内のすべて(あるいは指定数以上)のページからリンクを受けるページを新しい Center として抽出する．以下同様に，Center 集合からの新しい Fan 集合の抽出，Fan 集合からの新しい Center 集合の抽出を繰り返すことで，ページ同士の関連が強い Center 集合を作成し，これを Web コミュニティとする．初期の Fan の与え方として，Kumar らは HITS[7]の HUB を応用する方法を示している．

また，Flake らは，集合内のページ間のリンク数が集合以外のページとのリンク数より多いページの集合を Web コミュニティとし，最大流問題を解くことにより Web コミュニティを抽出できることを示している[8].

3.2 TF-IDF

TF-IDF[9]は文書内の単語につける重みの一種であり，以下の TF と IDF により決定される．

TF（Term Frequency）は，文書内におけるその単語の出現頻度を表す．TF が大きい単語は文書内に頻繁に出現し，そのような単語は文書内において重要，代表的な単語であると考えることができる．

IDF（Inverse Document Frequency）は，その単語を含む文書の頻度の逆数を表す．DF は文書頻度であり，その単語を含む文書の個数を表す．よって，DF が大きい単語は多くの文書に登場する一般的で重要度の低い単語と考えることができ，IDF が大きい単語が代表的な単語に適していると考えることができる．

tfidf 値は，以下の式(1)の様に TF と IDF により計算される．tfidf 値が高い単語ほど，その文書にとって重要な単語と考えることができる．ただし，n は文書内におけるその単語の出現回数，max（n）は文書内の全単語数，D は全文書数，d はその単語を含む文書数を表す．

$$\left.\begin{aligned} \mathbf{tfidf} &= \mathbf{tf} \times \mathbf{idf} \\ \mathbf{tf} &= \frac{\boldsymbol{n}}{\mathbf{max(n)}} \\ \mathbf{idf} &= \mathbf{log}\left(\frac{\mathbf{D}}{\mathbf{d}}\right) \end{aligned}\right\} \quad (1)$$

3.3 動画検索に関する研究

以下に，動画検索に関する研究を示す．中村らは，動画を再生時間軸に基づき印象分析し，動画の再生時間軸において喜びの度合いや悲しみの度合い，肯定度合，否定度合などを可視化させることで，印象の基づく動画検索を提案している[10].

中村らは，動画視聴サイトで推薦されるべき動画を「ユーザの直前の履歴に似ている動画」と「直前の履歴に似ていないがユーザの興味を引く，発見性のある動画」の 2 種類に分類し，動画視聴履歴データと動画間のメタデータ類似度を組み合わせた動画推薦を提案している[11].

江端らは，動画を視聴したユーザが付与することのできる唯一の情報がコメントであると考え，ユーザコメントに対し TF-IDF を用いて，あるユーザが視聴した動画コメントと，他の動画のコメントの類似度を計算し，関連動画を提示する手法を提案している[12].

平澤らは，ニコニコ動画で提供されているタグ機能に着目し，「もっと評価されるべき」タグの分析を行った．このタグを利用することで，あまり知られていないが多くの人が興味・関心のある動画を発見できることを確認した[13].

古尾らは，動画共有サイトを利用しているユーザ同士の繋がりから関連動画を得て，意外性のある動画推薦を提案している[14].

Web コミュニティや本研究の手法と同様にグラフや共引用(co-citation)の概念を用いた手法として，Baluja らによる視聴履歴から co-view を抽出する手法がある[15]．しかし，本手法は動画リストを用いておらずユーザの視聴履歴を元に動画を推薦することに主眼をおいた手法となっている．よって，動画リストを用いて個人の情報を用いない動画検索を提供する我々の手法とは，貢献の内容が異なっている．

4. 提案手法

動画共有サイトにおいて「共通の話題を持つ動画の集合」を動画コミュニティとする．本章において，Web コミュニティ抽出手法を単純に適用して動画コミュニティを抽出する手法(WC 手法)と，Web コミュニティ抽出手法と TF-IDF を併用して動画コミュニティを抽出する手法(WCTI 手法)の 2 つを提案する．

4.1 Web コミュニティ抽出手法を用いる動画コミュニティ (WC 手法)

まず，Web コミュニティ抽出手法のみを用いる WC 手法について述べる．WC 手法では表 1 の様に，Web コミュニティ抽出におけるリンク元ページ，リンク先ページ，リンクを，動画共有サイトにおけるものに置き換え，Web コミュニティ抽出手法を動画共有サイトに適用する手法である．

表 1 Web コミュニティ抽出手法を動画共有サイトに適用

Webコミュニティ抽出	動画コミュニティ抽出
Center（リンク先ページ）	動画
Fan（リンク元ページ）	動画リスト
Fan から Center へのリンク	動画リストによる動画の登録

本手法では，Fan である動画リスト群から多くの登録を受けている動画を Center 動画とし，Center 動画群に対して多くの登録を行っている動画リストを Fan 動画リストとする．

動画コミュニティの抽出は図 5 の手順に従いを行う．まず，初期 Center 動画の集合を用意する．そして，Center 動画集合内の動画を指定個数以上登録している動画リストを Fan 動画リストとして抽出する．続いて，Fan 動画リスト集合内の指定個数以上の動画リストに登録されている動画を Center 動画として抽出する．以下同様に，Center 動画集合からの Fan 動画リストの抽出，Fan 動画リスト集合からの Center 動画の抽出を収束する(Center と Fan に変化がなくなる)まで繰り返し，収束後の Center 動画集合を動画コミュニティとする．

4.2 Web コミュニティ抽出手法と TF-IDF を用いる動画コミュニティ (WCTI 手法)

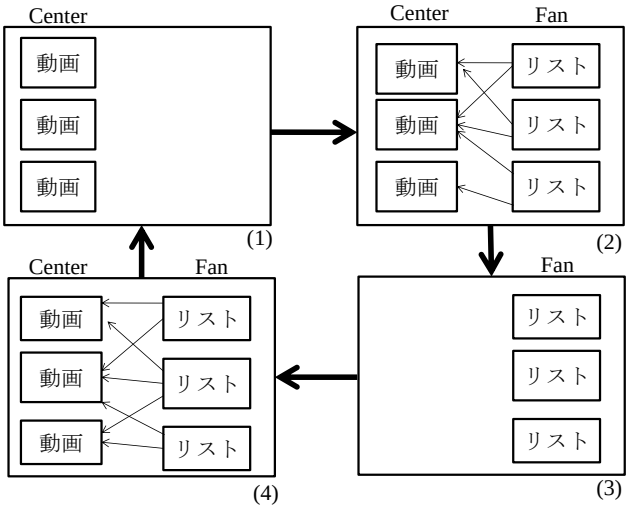

図 5 動画コミュニティ抽出法(WC 手法)

次に，Web コミュニティ抽出手法と TF-IDF を用いる動画コミュニティ抽出法(WCTI 手法)について述べる．

本手法では，検索語のタグが存在していることを前提に，検索語に関連する動画の動画コミュニティを抽出する．

前節の WC 手法同様に本手法でも表 1 の対応により，Web コミュニティ抽出におけるリンク元ページ，リンク先ページ，リンクを，動画共有サイトにおける動画リスト，動画，動画リストへの動画の登録で置き換え，Web コミュニティ抽出手法を適用する．

また，次項に示す方法で動画リスト内のタグに対して tfidf 値を定義し，Center 動画集合からの Fan 動画リストの抽出と Fan 動画リストの集合からの Center 動画の抽出の際にこの tfidf 値を用いて重み付けを行う．

4.2.1 動画リストとタグにおける TF-IDF

図 2 の様に，動画リスト内には 1 個以上の動画が登録されており，それぞれの動画にはタグが登録されている．本手法では，表 2 の対応により TF-IDF における文書，単語，文書内の全単語を動画共有サイトにおけるものに置き換え，TF-IDF を動画共有サイトに適用する．そして，3.2 節の式(1)を用いて動画リスト内のタグの tfidf 値を定義する．すなわち，n はその動画リスト内にそのタグが登場する回数，max(n)はその動画リスト内の全タグ数，D は全動画リスト数，d はそのタグを含む動画リストの数となる．

表 2 TF-IDF を動画共有サイトに適用

TF-IDF	動画共有サイトTF-IDF
文書	動画リスト
単語	動画のタグ
文書内の全単語	動画リスト内の全動画の全タグ

以下に，図 2 の例における，TF と IDF を示す．動画リスト X に存在するタグの数が 12 であるため，動画リスト X の全タグ数(max(n)相当)は 12 となる．タグ C は動画リスト X に 2 回登場するため，タグ C の動画リスト X における TF 値は 2/12 となる．タグ A とタグ F の動画リスト X における TF 値は 3/12, 1/12 となる．また，動画リストは 2 個であるため，全動画リスト数(D 相当)は 2 なる．そして，タグ F は動画リスト X のみに登場するためタグ F の DF 値は 1/2，IDF は log(2/1)となる．

4.2.2 TF-IDF を考慮した Center 集合から Fan の抽出

WCTI 手法では，前項の手法に従い各動画リスト内における検索語の tfidf 値を計算し，これが高い動画リストを検索語に適した動画リストとみなす．具体的には，動画リスト l の評価を以下の式(2)の fti(l)により行い，fti(l)が高い 100 件の動画リストを Fan とする．

$$\mathrm{fti}(l) = \mathrm{tfidf}^{n} \times mt \times f(l) \qquad (2)$$

ただし，$f(l)$は動画リスト l が含む Center 動画の数，mt はその動画リスト内の最高 tfidf 値，tfidf は動画リスト l における検索語の tfidf 値， tfidf^{n} はその n 乗である．n は tfidf の重みを表すチューニングパラメータで，1 より大きい数を想定している．1 より大きい理由は，mt と $f(l)$に対して tfidf の影響を相対的に大きくするためである．

mt(最高 tfidf 値)を用いている理由は，動画リスト内に高

い tfidf 値を持つタグが存在すれば，その動画リストは話題の一貫性が高く，存在しなければ話題の一貫性が低い動画リストであると期待できるからである．

前節の WC 手法と本節の WCTI 手法の違いは，前節の WC 手法は *f*(*l*)のみで評価しており，本節の WCTI 手法は *f*(*l*)と *mt*，tfidf の 10 乗の積で評価している点である．

4.2.3 TF-IDF を考慮した Fan 集合から Center の抽出

Fan 動画リスト集合からの Center 動画の抽出においては，以下の式(3) の cti(*v*)により動画 *v* を評価し，cti(*v*)が高い 50 件の動画を Center とする．

$$\mathrm{cti}(v) = HasTag(v,t) + \sum_{l \in L} \mathrm{fti}(l) \qquad (3)$$

ただし，*L* は，「Center 動画を含んでいる動画リスト」の集合，*HasTag*(*v*,*t*)の値は動画 *v* が検索語である *t* をタグに持てば 1，持たなければ 0 である．

$\sum_{l \in L} \mathrm{fti}(l)$は，動画 *v* が Fan 動画リストから含まれるごとに評価値 fti(*l*)を与えられ，その評価値 fti(*l*)は Center への関連度と検索語への関連度を加味したものとなっている．$\sum_{l \in L} \mathrm{fti}(l)$は 1 と比べて十分に小さいため，実質的にはタグの有無(*HasTag* が 1 であるか 0 であるか)により動画がクラス分けされ，同一クラス内における優先順位付けに$\sum_{l \in L} \mathrm{fti}(l)$が使用されることとなる．

4.2.4 WCTI 手法による動画コミュニティ抽出手順

WCTI 手法における動画コミュニティ抽出手順を図 6 および以下の(1)～(4)に示す．

（1） 共通の主題を持った動画を 10 件選択し，それを初期の Center 集合とする．初期 Center 動画の選定方法は動画共有サイトの実装に依存し，具体的な手法は次章にて述べる．

（2） 第 4.2.2 項の手順に従い，動画リストを fti(*l*)を用いて評価する．そして，値が高い動画リスト 100 件を Fan 動画リスト集合とする．

（3） 第 4.2.3 項の手順に従い，動画を cti(*v*)を用いて評価する．そして，値が高い動画 50 件を Center 動画集合とする．

（4） 収束をする(Center と Fan に変化がなくなる)まで，上記の(2)と(3)を繰り返す．

以上により得られた Center 集合を動画コミュニティとする．

5. 評価

本章では，動画共有サイトで提供されている検索機能，Web 検索エンジン，提案手法(WC 手法)，提案手法(WCTI 手法)のそれぞれによる検索結果の比較を行う．

動画共有サイトにより提供されている検索手法の検索結果としては，キーワード検索結果を再生回数順あるいは動画リスト登録回数順に並び替えて上位 50 件を検索結果としたもの，検索語をタグに含む動画群を再生回数順あるいは動画リスト登録回数順に並び替え上位 50 件を検索結果としたもの，の 4 通りを用いた．また，Web 検索エンジンは検索範囲を当該動画共有サイトのみに指定し単語検索を行った上位 50 件を検索結果とした．両提案手法では，抽出された動画コミュニティ内の動画の上位 50 件を検索結果とし，提案手法の tfidf の重み *n* には 10 を用いた．両提案手法の初期 Center 動画の集合としては，動画共有サイトにより提供されているタグ検索の結果を動画リスト登録回数順に並び替えた上位 10 件を選択したものを用いた．動画共有サイトにはニコニコ動画を用い，抽出は 2012 年 7 月 1 日から 2013 年 4 月 7 日にニコニコ動画より収集した 1,550,738 件の動画と，127,955 件の動画リストを用いて行った．

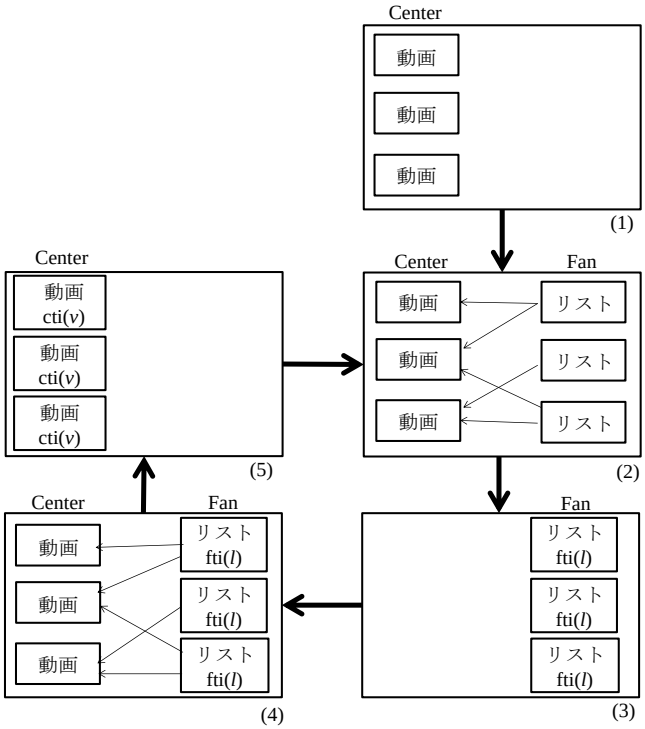

図 6 TF-IDF 法を用いる動画コミュニティ抽出法 (WCTI 手法)

検索結果の評価は 6 人の被験者が検索結果に含まれる各動画を再生，閲覧し主観により次の 3 段階の評価(A, B, C)に分類した． 評価者には著者は含まれていない．

(A 評価)：検索語と深い関係がある動画
(B 評価)：検索語と関連があるが，関係が深くない動画
(C 評価)：検索語と無関係の動画

評価者に対しては，全手法の検索結果に含まれる全動画の一覧のみが与えられ，どの動画がどの検索手法による検索結果であるかを評価者が特定できない状況で評価を行った．検索語を「世界遺産」としたとき，「チャーハン」としたとき，「MTG」としたとき，政治家の名前(以下「政治家」)としたときの評価結果を表 3 から表 6 に示す．それぞれの表は一人の評価者により評価結果を表しており，表の「合計」は検索語を 3 段階評価に分類したときの(A 評価) [+1 点]，(B 評価)[±0 点]，(C 評価)[-1 点]の合計値である．

表 3 (a) 検索語「世界遺産」の評価(1)

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	0	14	36	−36
キーワード検索＋動画リスト登録数が多い順	3	13	34	−31
タグ検索＋再生数が多い順	9	21	20	−11
タグ検索＋動画リスト登録数が多い順	12	19	19	−7
Web検索エンジン(ニコニコ動画のみを対象とする)	20	17	13	7
WC手法	0	11	39	−39
WCTI手法	38	10	2	36

表 3 (b) 検索語「世界遺産」の評価(2)

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	2	3	45	−43
キーワード検索＋動画リスト登録数が多い順	4	4	42	−38
タグ検索＋再生数が多い順	15	8	27	−12
タグ検索＋動画リスト登録数が多い順	14	9	27	−13
Web検索エンジン(ニコニコ動画のみを対象とする)	34	5	11	23
WC手法	2	8	40	−38
WCTI手法	40	5	5	35

表 3© 検索語「世界遺産」の評価(3)

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	4	14	32	−28
キーワード検索＋動画リスト登録数が多い順	7	13	30	−23
タグ検索＋再生数が多い順	14	21	15	−1
タグ検索＋動画リスト登録数が多い順	19	16	15	4
Web検索エンジン(ニコニコ動画のみを対象とする)	27	11	12	15
WC手法	4	18	28	−24
WCTI手法	44	6	0	44

表 4 (a) 検索語「チャーハン」の評価(1)

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	5	4	41	−36
キーワード検索＋動画リスト登録数が多い順	1	6	43	−42
タグ検索＋再生数が多い順	28	7	15	13
タグ検索＋動画リスト登録数が多い順	22	10	18	4
Web検索エンジン(ニコニコ動画のみを対象とする)	23	2	25	−2
WC手法	0	0	50	−50
WCTI手法	40	5	5	35

表 4 (b) 検索語「チャーハン」の評価(2)

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	5	0	45	−40
キーワード検索＋動画リスト登録数が多い順	1	0	49	−48
タグ検索＋再生数が多い順	26	5	19	7
タグ検索＋動画リスト登録数が多い順	24	3	23	1
Web検索エンジン(ニコニコ動画のみを対象とする)	25	1	24	1
WC手法	0	1	49	−49
WCTI手法	43	1	6	37

表 4 (c) 検索語「チャーハン」の評価(3)

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	4	3	43	−39
キーワード検索＋動画リスト登録数が多い順	2	1	47	−45
タグ検索＋再生数が多い順	25	9	16	9
タグ検索＋動画リスト登録数が多い順	22	8	20	2
Web検索エンジン(ニコニコ動画のみを対象とする)	22	1	27	−5
WC手法	1	2	47	−46
WCTI手法	40	5	5	35

表 5 (a) 検索語「MTG」の評価(1)

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	8	15	27	−19
キーワード検索＋動画リスト登録数が多い順	4	13	33	−29
タグ検索＋再生数が多い順	8	15	27	−19
タグ検索＋動画リスト登録数が多い順	4	14	32	−28
Web検索エンジン(ニコニコ動画のみを対象とする)	41	7	2	39
WC手法	0	0	50	−50
WCTI手法	45	1	4	41

表 5 (b) 検索語「MTG」の評価(2)

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	16	30	4	12
キーワード検索＋動画リスト登録数が多い順	8	31	11	−3
タグ検索＋再生数が多い順	16	32	2	14
タグ検索＋動画リスト登録数が多い順	9	36	5	4
Web検索エンジン(ニコニコ動画のみを対象とする)	42	7	1	41
WC手法	0	3	47	−47
WCTI手法	46	3	1	45

表 5 (c) 検索語「MTG」の評価(3)

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	7	35	8	−1
キーワード検索＋動画リスト登録数が多い順	3	31	16	−13
タグ検索＋再生数が多い順	7	37	6	1
タグ検索＋動画リスト登録数が多い順	4	35	11	−7
Web検索エンジン(ニコニコ動画のみを対象とする)	32	16	2	30
WC手法	0	1	49	−49
WCTI手法	46	0	4	42

表 6 (a) 検索語「政治家 A」の評価(1)

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	22	20	8	14
キーワード検索＋動画リスト登録数が多い順	23	20	7	16
タグ検索＋再生数が多い順	24	19	7	17
タグ検索＋動画リスト登録数が多い順	25	20	5	20
Web検索エンジン(ニコニコ動画のみを対象とする)	40	8	2	38
WC手法	25	12	13	12
WCTI手法	50	0	0	50

表 6 (b) 検索語「政治家 A」の評価(2)

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	17	16	17	0
キーワード検索＋動画リスト登録数が多い順	13	16	21	−8
タグ検索＋再生数が多い順	17	16	17	0
タグ検索＋動画リスト登録数が多い順	12	16	22	−10
Web検索エンジン(ニコニコ動画のみを対象とする)	31	12	7	24
WC手法	12	17	21	−9
WCTI手法	50	0	0	50

表 6 (c) 検索語「政治家 A」の評価(3)

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	12	29	9	3
キーワード検索＋動画リスト登録数が多い順	16	28	6	10
タグ検索＋再生数が多い順	12	30	8	4
タグ検索＋動画リスト登録数が多い順	17	29	4	13
Web検索エンジン(ニコニコ動画のみを対象とする)	33	14	3	30
WC手法	12	27	11	1
WCTI手法	50	0	0	50

すべての評価結果おいて，Web コミュニティ抽出手法と TF-IDF を併用する WCTI 手法が最も(A 評価)が多く，また(C 評価)が少なくなり，本提案手法が有効であることが分かった．また，単純に Web コミュニティ抽出手法を用いたのみの WC 手法では動画コミュニティを抽出することが可能であったとしても，それをそのまま検索に応用することはできないことも分かった．

WC 手法では，初期の Center 集合として検索語と適合度が高い動画群を与えても，Fan の抽出と Center の抽出を繰り返すに従い検索語を含むより抽象的な概念の動画の集合と変わって行ってしまった．たとえば，初期 Center に「チャーハン」の動画群を与えた場合は，「チャーハン」より抽象度が高い「料理」に関する動画の集合に収束した．

WC 手法と WCTI 手法の比較より，動画リストには話題の選定に有効なものもあるが，有効でないものも含まれており，検索にはこれらの区別が重要であると考えられる．

次に，WCTI 手法における tfidf の値の重み n に関する評価を行う．WCTI 手法では，mt と $f(l)$に対して tfidf の影響を大きくするために，式(2)の様に tfidf を n 乗して fti(l)を求めている．tfidf の重み n を変化させて評価を行った．

評価結果を表 7 から表 10 に示す．表より，n を大きくしないと，良い結果が得られないことが分かるが，一定以上(5 以上)にすることにより十分に良い性能が得られることが分かる．また，n を一定以上大きくすれば，性能は n に依存せず，n の最適化が不要であることが分かる．

表 7 (a) 検索語「世界遺産」の$\boldsymbol{tfidf^n}$の評価(1)

	A(+1)	B(±0)	C(−1)	合計
式(4)を用いるWCTI手法(n=1)	23	20	7	16
式(4)を用いるWCTI手法(n=2)	26	18	6	20
式(4)を用いるWCTI手法(n=3)	38	10	2	36
式(4)を用いるWCTI手法(n=4)	38	10	2	36
式(4)を用いるWCTI手法(n=5)	38	10	2	36
式(4)を用いるWCTI手法(n=6)	38	10	2	36
式(4)を用いるWCTI手法(n=7)	38	10	2	36
式(4)を用いるWCTI手法(n=8)	38	10	2	36
式(4)を用いるWCTI手法(n=9)	38	10	2	36
式(4)を用いるWCTI手法(n=10)	38	10	2	36

表 7 (b) 検索語「世界遺産」の$\boldsymbol{tfidf^n}$の評価(2)

	A(+1)	B(±0)	C(−1)	合計
式(4)を用いるWCTI手法(n=1)	32	7	11	21
式(4)を用いるWCTI手法(n=2)	30	10	10	20
式(4)を用いるWCTI手法(n=3)	40	5	5	35
式(4)を用いるWCTI手法(n=4)	40	5	5	35
式(4)を用いるWCTI手法(n=5)	40	5	5	35
式(4)を用いるWCTI手法(n=6)	40	5	5	35
式(4)を用いるWCTI手法(n=7)	40	5	5	35
式(4)を用いるWCTI手法(n=8)	40	5	5	35
式(4)を用いるWCTI手法(n=9)	40	5	5	35
式(4)を用いるWCTI手法(n=10)	40	5	5	35

表 7 (c) 検索語「世界遺産」の$\boldsymbol{tfidf^n}$の評価(3)

	A(+1)	B(±0)	C(−1)	合計
式(4)を用いるWCTI手法(n=1)	30	15	5	25
式(4)を用いるWCTI手法(n=2)	36	11	3	33
式(4)を用いるWCTI手法(n=3)	44	6	0	44
式(4)を用いるWCTI手法(n=4)	44	6	0	44
式(4)を用いるWCTI手法(n=5)	44	6	0	44
式(4)を用いるWCTI手法(n=6)	44	6	0	44
式(4)を用いるWCTI手法(n=7)	44	6	0	44
式(4)を用いるWCTI手法(n=8)	44	6	0	44
式(4)を用いるWCTI手法(n=9)	44	6	0	44
式(4)を用いるWCTI手法(n=10)	44	6	0	44

表 8 (a) 検索語「チャーハン」の$\boldsymbol{tfidf^n}$の評価(1)

	A(+1)	B(±0)	C(−1)	合計
式(4)を用いるWCTI手法(n=1)	38	6	6	32
式(4)を用いるWCTI手法(n=2)	34	6	10	24
式(4)を用いるWCTI手法(n=3)	34	5	11	23
式(4)を用いるWCTI手法(n=4)	33	5	12	21
式(4)を用いるWCTI手法(n=5)	33	5	12	21
式(4)を用いるWCTI手法(n=6)	33	5	12	21
式(4)を用いるWCTI手法(n=7)	40	5	5	35
式(4)を用いるWCTI手法(n=8)	40	5	5	35
式(4)を用いるWCTI手法(n=9)	40	5	5	35
式(4)を用いるWCTI手法(n=10)	40	5	5	35

表 8 (b) 検索語「チャーハン」の$\boldsymbol{tfidf^n}$の評価(2)

	A(+1)	B(±0)	C(−1)	合計
式(4)を用いるWCTI手法(n=1)	41	1	8	33
式(4)を用いるWCTI手法(n=2)	37	1	12	25
式(4)を用いるWCTI手法(n=3)	36	1	13	23
式(4)を用いるWCTI手法(n=4)	35	1	14	21
式(4)を用いるWCTI手法(n=5)	35	1	14	21
式(4)を用いるWCTI手法(n=6)	35	1	14	21
式(4)を用いるWCTI手法(n=7)	43	1	6	37
式(4)を用いるWCTI手法(n=8)	43	1	6	37
式(4)を用いるWCTI手法(n=9)	43	1	6	37
式(4)を用いるWCTI手法(n=10)	43	1	6	37

表 8 (c) 検索語「チャーハン」の$\boldsymbol{tfidf^n}$の評価(3)

	A(+1)	B(±0)	C(−1)	合計
式(4)を用いるWCTI手法(n=1)	38	6	6	32
式(4)を用いるWCTI手法(n=2)	35	5	10	25
式(4)を用いるWCTI手法(n=3)	34	5	11	23
式(4)を用いるWCTI手法(n=4)	33	5	12	21
式(4)を用いるWCTI手法(n=5)	33	5	12	21
式(4)を用いるWCTI手法(n=6)	33	5	12	21
式(4)を用いるWCTI手法(n=7)	40	5	5	35
式(4)を用いるWCTI手法(n=8)	40	5	5	35
式(4)を用いるWCTI手法(n=9)	40	5	5	35
式(4)を用いるWCTI手法(n=10)	40	5	5	35

表 9 (a) 検索語「MTG」の$\boldsymbol{tfidf^n}$の評価(1)

	A(+1)	B(±0)	C(−1)	合計
式(4)を用いるWCTI手法(n=1)	42	3	5	37
式(4)を用いるWCTI手法(n=2)	46	0	4	42
式(4)を用いるWCTI手法(n=3)	45	1	4	41
式(4)を用いるWCTI手法(n=4)	45	1	4	41
式(4)を用いるWCTI手法(n=5)	45	1	4	41
式(4)を用いるWCTI手法(n=6)	45	1	4	41
式(4)を用いるWCTI手法(n=7)	45	1	4	41
式(4)を用いるWCTI手法(n=8)	45	1	4	41
式(4)を用いるWCTI手法(n=9)	45	1	4	41
式(4)を用いるWCTI手法(n=10)	45	1	4	41

表 9 (b) 検索語「MTG」の$\boldsymbol{tfidf^n}$の評価(2)

	A(+1)	B(±0)	C(−1)	合計
式(4)を用いるWCTI手法(n=1)	43	6	1	42
式(4)を用いるWCTI手法(n=2)	47	2	1	46
式(4)を用いるWCTI手法(n=3)	46	3	1	45
式(4)を用いるWCTI手法(n=4)	46	3	1	45
式(4)を用いるWCTI手法(n=5)	46	3	1	45
式(4)を用いるWCTI手法(n=6)	46	3	1	45
式(4)を用いるWCTI手法(n=7)	46	3	1	45
式(4)を用いるWCTI手法(n=8)	46	3	1	45
式(4)を用いるWCTI手法(n=9)	46	3	1	45
式(4)を用いるWCTI手法(n=10)	46	3	1	45

表 9 (c) 検索語「MTG」の$\boldsymbol{tfidf^n}$の評価(3)

	A(+1)	B(±0)	C(−1)	合計
式(4)を用いるWCTI手法(n=1)	43	2	5	38
式(4)を用いるWCTI手法(n=2)	46	0	4	42
式(4)を用いるWCTI手法(n=3)	46	0	4	42
式(4)を用いるWCTI手法(n=4)	46	0	4	42
式(4)を用いるWCTI手法(n=5)	46	0	4	42
式(4)を用いるWCTI手法(n=6)	46	0	4	42
式(4)を用いるWCTI手法(n=7)	46	0	4	42
式(4)を用いるWCTI手法(n=8)	46	0	4	42
式(4)を用いるWCTI手法(n=9)	46	0	4	42
式(4)を用いるWCTI手法(n=10)	46	0	4	42

表 10 (a) 検索語「政治家 A」の$\boldsymbol{tfidf^n}$の評価(1)

	A(+1)	B(±0)	C(−1)	合計
式(4)を用いるWCTI手法(n=1)	29	1	20	9
式(4)を用いるWCTI手法(n=2)	37	1	12	25
式(4)を用いるWCTI手法(n=3)	48	2	0	48
式(4)を用いるWCTI手法(n=4)	49	1	0	49
式(4)を用いるWCTI手法(n=5)	50	0	0	50
式(4)を用いるWCTI手法(n=6)	50	0	0	50
式(4)を用いるWCTI手法(n=7)	50	0	0	50
式(4)を用いるWCTI手法(n=8)	50	0	0	50
式(4)を用いるWCTI手法(n=9)	50	0	0	50
式(4)を用いるWCTI手法(n=10)	50	0	0	50

表 10 (b) 検索語「政治家 A」の$tfidf^n$の評価**(2)**

	A(+1)	B(±0)	C(−1)	合計
式(4)を用いるWCTI手法(n=1)	26	2	22	4
式(4)を用いるWCTI手法(n=2)	36	2	12	24
式(4)を用いるWCTI手法(n=3)	48	2	0	48
式(4)を用いるWCTI手法(n=4)	49	1	0	49
式(4)を用いるWCTI手法(n=5)	50	0	0	50
式(4)を用いるWCTI手法(n=6)	50	0	0	50
式(4)を用いるWCTI手法(n=7)	50	0	0	50
式(4)を用いるWCTI手法(n=8)	50	0	0	50
式(4)を用いるWCTI手法(n=9)	50	0	0	50
式(4)を用いるWCTI手法(n=10)	50	0	0	50

表 10 (c) 検索語「政治家 A」の$tfidf^n$の評価**(3)**

	A(+1)	B(±0)	C(−1)	合計
式(4)を用いるWCTI手法(n=1)	28	2	20	8
式(4)を用いるWCTI手法(n=2)	36	2	12	24
式(4)を用いるWCTI手法(n=3)	48	2	0	48
式(4)を用いるWCTI手法(n=4)	49	1	0	49
式(4)を用いるWCTI手法(n=5)	50	0	0	50
式(4)を用いるWCTI手法(n=6)	50	0	0	50
式(4)を用いるWCTI手法(n=7)	50	0	0	50
式(4)を用いるWCTI手法(n=8)	50	0	0	50
式(4)を用いるWCTI手法(n=9)	50	0	0	50
式(4)を用いるWCTI手法(n=10)	50	0	0	50

6. おわりに

本稿では，動画コミュニティ抽出法と TF-IDF を使用した動画検索手法を提案した．提案手法と他の検索手法の検索結果を評価した結果，提案手法は他の検索手法に比べて検索語と関連の高い動画をより多く抽出可能であることが確認され，有効性が示された．

今後は，さらに多くの検索語による評価をし，さらに精度を向上させるための方法を考察する予定である．

謝辞

検索結果の評価のために，非常に多くの動画の閲覧を行った評価者の人たちに感謝の意を表する．

参考文献

[1] 動画サイトの利用実態調査検討委員会 -報告書- http://www.riaj.or.jp/release/2011/pdf/20110808_2report.pdf

[2] 西友規,山口実靖, “動画共有サイトにおける動画リストを用いた動画検索”,情報処理学会研究会報告,データベース・システム研究会報告 2012-DBS-156(10),1-6,2012.

[3] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins,“Trawling the Web for emerging cyber communities,” In Proc. of the 8th international conference on World Wide Web, pp. 1481 - 1493, 1999.

[4] Polepalli Krishna Reddy, and Masaru Kitsuregawa, “An approach to relate the web communities through bipartitegraphs,” Proc. of the 2nd International Conference on Web Information Systems Engineering, 2001.

[5] 村田 剛志, “参照の共起性に基づく Web コミュニティの発見”, 人工知能学会論文誌, Vol. 16, No. 3, pp. 316-323, 2001.

[6] Naoyuki Saida, Akira Umezawa, and Hayato Yamana, “PlusDBG: Web Community Extraction Scheme Improving Both Precision and Pseudo-Recall”, In Proc. of the 7th Asia-Pacific Web Conference, 2005.

[7] Jon Michael Kleinberg, “Authoritative sources in a hyperlinked environment,” Journal of the ACM (JACM), Volume 46 Issue 5, pp. 604 - 632, 1999.

[8] Gary William Flake, Steve Lawrence, and Clyde. Lee Giles, “Efficient identification of Web communities,” In Proc. of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 150 - 160, 2000.

[9] Gerard Salton, and Christopher Buckley, “Term-weighting approaches in automatic text retrieval,” Inf. Process. and Management, vol.24, no.5, pp. 513 - 523, 1988.

[10] 中村聡史,田中克己“印象に基づく動画検索”,情報処理学会告.HCI, ヒューマンコンピュータインタラクション研究会報告 2009(5),77-84,2009.

[11] 中村智浩,山名早人, “動画視聴サイトにおける発見性を重視した動画推薦手法の提案”,DEIM2010 A3-1,2010.

[12] 江端佑介,川村秀憲,鈴木恵二, “ユーザコメントの tf-idf 法による分析を用いたインタラクティブな関連動画の提示”,電子情報通信学会技術研究報告.AI,人工知能と知識処理 109(439),7-10,2010.

[13] 平澤真大,小川佑樹,諏訪博彦,太田敏澄, “ニコニコ動画のログデータを用いたソーシャルノベルティのある動画の発見に関する研究”,情報処理学会研究会報告,データベース・システム研究会報告 2011-DBS-153(13),1-8,2011.

[14] 古尾透,太田学,“ユーザの繋がりを用いた意外性のある動画推薦システム” ,DEIM2012 B4-1,2012.

[15] Shumeet Baluja, Rohan Seth, D. Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran and Mohamed Aly, "Video Suggestion and Discovery for YouTube: Taking Random Walks through the View Graph", Proc. of the 17th international conference on World Wide Web, pp. 895 - 904, 2008.