Vol.2014-DBS-159 No.1
Vol.2014-IFAT-115 No.1
2014/8/1

リンク構造に基づくページ検索手法を用いた動画検索に関する一考察

西友規† 山口実靖† 小林亜樹†

動画共有サービスが普及し，多くの動画が動画共有サイトで共有されている．しかし，既存の動画キーワード検索手法の精度は必ずしも十分とは言えず，動画のキーワード検索の精度向上は重要な課題の一つと考えることができる．動画共有サイトでユーザが公開している動画リストと公開動画リストに登録されている動画の関係は有向グラフで表すことができる．本稿では，Web ページ間のリンク構造を解析し，ランキングを行う HITS アルゴリズムに着目し，これを応用したキーワードとの関連度を考慮した動画検索についての考察を行う．

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，動画共有サイトで提供されている動画のキーワード検索結果は再生回数順などの人気順で提供されることが多く，必ずしも検索語との関連を考慮した検索とはなっていない．よって，動画共有サイトにおける単語による動画検索の精度(適合率)の向上は重要な課題の一つと考えることができる．

動画共有サイトでユーザが公開している動画リストとその動画リストに登録されている動画の関係は有向グラフで表すことができ，既存の Web ページ間のリンク構造を解析する手法を動画共有サイトに適用することが可能であると予想できる．

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，既存研究である動画コミュニティの抽出手法[2,3,4]，HITS アルゴリズム[5]，動画検索に関する研究を紹介する．そして，Web ページ間のリンク構造を解析しランキングを行う HITS アルゴリズムに着目し，これを応用し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 動画コミュニティ抽出手法

3.1.1 WC 手法

WC 手法[2]は，Web コミュニティ抽出手法[6]を動画共有サイトに適用した手法である．Web コミュニティ抽出手法における Center(リンク元ページ)，Fan(リンク先ページ)，Fan から Center へのリンクを，動画共有サイトにおける動

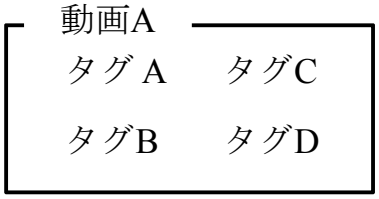

図 1 動画とタグ

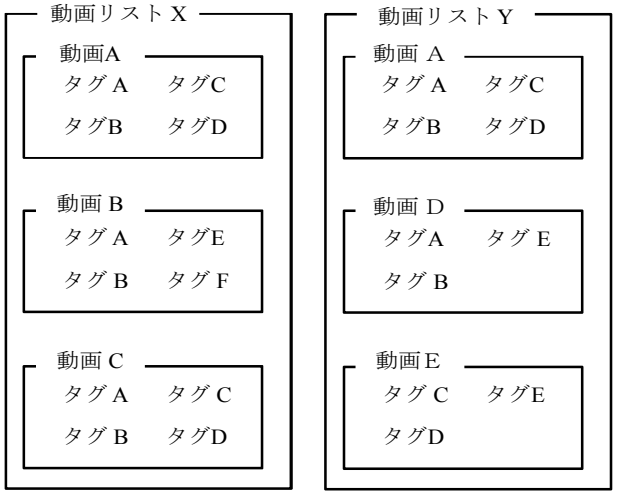

図 2 動画リスト

† 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

Vol.2014-DBS-159 No.1
Vol.2014-IFAT-115 No.1
2014/8/1

画, 動画リスト, 動画リストによる動画の登録に置き換え, 動画共有サイトに適用している.

WC 手法における動画コミュニティ抽出手順を図 3 および以下の(1)～(4)に示す.

（1）共通の話題を持った動画を指定数選択し，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する.

（2）Center 動画集合を登録している全動画リストを抽出し，Center 動画集合内の動画をより多く登録している上位 x 件の動画リストを Fan 動画リスト集合とする.

（3）Fan 動画リスト集合に登録されている全動画を抽出し，Fan 動画リスト集合内の動画リストからより多く登録されている上位 x 件の動画を 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.

3.1.2 WCTI 手法

WCTI 手法[3]は, WC 手法と TF-IDF を併用した手法である．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動画リスト，動画のタグ，動画リスト内の全動画の全タグに置き換え, 動画共有サイトに TF-IDF を適用している.

WCTI 手法では，WC 手法における Center 動画集合からの Fan の抽出の際に，動画リスト l を以下の fti(l)を用いて評価する．そして，以下の fti(l)値が高い動画リスト x 件を Fan 動画リスト集合とする.

$$\mathrm{fti}(l) = \mathrm{tfidf}^{10} \times mt \times f(l) \qquad (1)$$

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^{10}はその 10 乗である.

また，Center 動画集合からの Fan の抽出の際には，動画 v を以下の cti(v)を用いて評価する．そして，cti(v)値が高い動画 x 件を Center 動画集合とする.

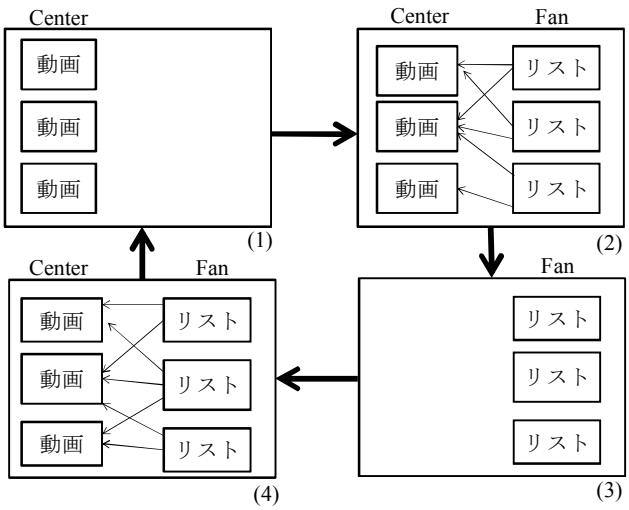

図 3 動画コミュニティ抽出手順

$$\mathrm{cti}(v) = HasTag(v,t) + \sum_{l \in L} \mathrm{fti}(l) \qquad (2)$$

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)を与えられる.

これら以外は，WC 手法と同一の手順を用いて動画コミュニティの抽出を行う.

3.1.3 WCTIZ 手法

WCTIZ 手法[3]は，WCTI 手法に Zipf の法則を用いた手法である. WCTIZ 手法では動画リスト内のタグの出現頻度が Zipf の法則に従うという仮説を立て Center 動画集合からの Fan 動画リストの際に，Zipf の法則を用いた動画リストの話題一貫性の強弱の判定を行い話題一貫性の低い動画リストを除外している.

具体的には，平均的な動画リストにおいてはタグ t の出現頻度 TF(t)と, その頻度順位 TFRank(t)の積は一定であり, 話題一貫性の低い動画リストにおいては t と TF(t)×TFRank(t)のグラフが右下がりとなるとの仮定を立て, 右下がりとなる動画リストを Fan から除外する.

それ以外は，WCTI 手法と同一の手順を用いて動画コミュニティの抽出を行う.

3.1.4 WCTI 手法と WCTIZ 手法を改良

WCTI 手法と WCTIZ 手法において Center 動画集合からの Fan 集合の抽出の際に, Center 動画の評価値である cti(v)値を考慮することにより，Center 動画の選定を改善している[4]. 具体的には，Center 動画集合からの Fan 集合の抽出の際に，以下の式(3)を用いることで，Center 動画の評価値 cti(v)を考慮する.

$$\mathrm{fti}(l) = \mathrm{tfidf}^{10} \times mt \times \sum_{v \in V} \mathrm{cti}(v) \qquad (3)$$

V は，「動画リストが登録している Center 動画」の集合である．ただし，初期 Center 動画集合からの Fan 動画リストの抽出のときのみ cti(v)値が存在しないため全ての cti(v)値を 1 とする.

WCTI 手法や WCTIZ 手法との違いは，両手法では Fan の評価の際にどの Center 動画に対して登録を行っても Fan 動画リストが得る評価値が一定であったが，改善手法では評価値の高い Center 動画を登録するほど Fan 動画リストはより高い評価値を得られる点である.

3.2 HITS

HITS アルゴリズム[5]は Kleinberg が提案した手法で, Web ページ間のリンク構造を解析することで Web ページのランキングを行う．HITS アルゴリズムでは Web ページに対し authority と hub の 2 つの尺度を与えて Web ページを評価する．authority とはある特定の話題に関する情報を多く持つ情報源となる Web ページであり，情報源の質が高い authority はより多くの hub からリンクされる．hub とは情

報源となる Web ページへのリンクを多く持つリンク集となる Web ページであり，リンク集としての質が高い hub は情報源の質が高い authority をより多くリンクしている．

このように authority と hub は相互依存の関係にあり，authority と hub の評価値は反復計算によって求められる．authority の評価値をx_i，hub の評価値をy_iとし，Web ページ i から Web ページ j へのリンクを$p_i \to p_j$とすると，x_iとy_iはそれぞれ式(4)，式(5)で求められる．また，式(6)と式(7)により正規化される．

$$x_i = \sum_{p_j \to p_i} y_j \tag{4}$$

$$y_i = \sum_{p_i \to p_j} x_j \tag{5}$$

$$x_i = \frac{x_i}{\sqrt{\sum_j^n x_j^2}} \tag{6}$$

$$y_i = \frac{y_i}{\sqrt{\sum_j^n y_j^2}} \tag{7}$$

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[7]．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[8]．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[9]．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[10]．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[11]．これらの手法はユーザの履歴情報を用いて，動画を推薦する手法であり，単純な単語検索の精度向上を目指す我々の手法とは目的が異なっている．

Web コミュニティやグラフの共引用(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[12]．しかし，当該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

本章では, HITS アルゴリズムを動画共有サイトに用いる動画検索手法，それを発展させた HITS アルゴリズムと動画再生数を考慮した動画検索手法および HITS アルゴリズムと TF-IDF を併用した動画検索手法について述べる．

4.1 HITS アルゴリズムを用いる動画検索手法(nHITS 手法)

HITS アルゴリズムを単純に適用するナイーブな手法について述べる．本手法を nHITS 手法と呼ぶ．nHITS 手法では，HITS アルゴリズムにおける「authority」を「動画」，「hub」を「公開動画リスト」，「ページからページへのリンク」を「動画リストによる動画の登録」に置き換え HITS アルゴリズムを動画共有サイトに適用する．nHITS 手法の手順を以下に示す．

(1) 検索語を動画共有サイトの検索システムに与え，r 件の動画を収集し，rootset 集合とする．具体的な手法は 5 章にて述べる．

(2) rootset 集合に含まれる動画を登録している公開動画リストを全て抽出し, rootset 集合に追加して baseset 集合とする．baseset 集合内のリンク構造を抽出し，二部グラフを作成し，初期値として各動画に authority の評価値$x_i = 1$，各公開動画リストに hub の評価値y_i =1 を与える．

(3) 式(4)を用いて baseset 集合内の動画に authority の評価値x_iを求め，式(6)により正規化する．そして，評価値x_iの値が高い順に並べ替えたものを動画集合とする．

(4) 式(5)を用いて baseset 集合内の公開動画リストに hub の評価値を求め，式(7)により正規化する．そして，評価値y_iの値が高い順に並べ替えたものを公開動画リスト集合とする．

(5) 収束する(動画集合と公開動画リスト集合内の出現順に変化がなくなる)まで上記の(3)と(4)を繰り返す．

HITS では各 Web ページは authority と hub の両方の評価値を持つが nHITS では動画は authority のみ，動画リストは hub のみの評価値を持つ．

4.2 動画再生数を考慮した HITS アルゴリズムを用いる動画検索手法(vaHITS 手法，vhHITS 手法)

動画再生数を考慮して HITS アルゴリズムを用いる手法について述べる．authority(動画)の評価に動画再生数を考慮した手法を vaHITS 手法，hub(公開動画リスト)の評価に動画再生数を考慮した手法を vhHITS 手法と呼ぶ．4.1 節と同様に HITS アルゴリズムを動画共有サイトに適用する．動画の再生数が多い動画は動画共有サイトを利用するユーザの多くが視聴する動画であるため人気のある動画であり，高い評価の authority であると期待できる．

動画の再生数を v とすると，vaHITS 手法では動画の評価を求める際に式(4)ではなく式(8)を用いることで動画再生

数が多い動画はより多くの評価が与えられる仕組みとする．それ以外は，nHITS 手法と同一の手順を用いる．

$$x_i = \sum_{p_j \to p_i} y_j \times v_i \tag{8}$$

vhHITS 手法では公開動画リストの評価を求める際に式(5)ではなく式(9)を用いることで動画再生数の多い動画を登録しているとより多くの評価が与えられる仕組みとする．それ以外は，nHITS 手法と同一の手順を用いる．

$$y_i = \sum_{p_i \to p_j} (x_j \times v_j) \tag{9}$$

4.3 HITS アルゴリズムと TF-IDF を用いる動画検索手法 (tiHITS 手法)

HITS アルゴリズムと TF-IDF を併用した手法について述べる．本手法を tiHITS 手法と呼ぶ．4.1 節の nHITS 手法同様に，HITS アルゴリズムを動画共有サイトに適用する．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できる．そこで，tiHITS 手法では 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動画リスト，動画のタグ，動画リスト内の全動画の全タグに置き換え，TF-IDF を動画共有サイトに適用し，公開動画リスト内の各タグに tfidf 値を定義する．

tiHITS 手法では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．そして，動画の評価を式(4)ではなく式(10)により行い，検索語に関係のある動画リストから登録されている動画はより多くの評価が与えられるものとする．それ以外は，nHITS 手法と同一の手順を用いる．

$$x_i = \sum_{p_j \to p_i} (y_j \times \mathrm{tfidf}_j) \tag{10}$$

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(nHITS 手法，vaHITS 手法，vhHITS 手法，tiHITS 手法)のそれぞれによる検索結果の入力キーワードとの関連度を評価する．

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．これらの検索手法は検索語と関連の高い動画を抽出する手法ではないが，動画共有サイトにはこれら以外の手法が用意されていないため参考のためにこれらの検索結果との比較を掲載する．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．

提案手法では，抽出された動画集合内の動画の上位 50 件を検索結果とした．また，提案手法の rootset 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 200 件まで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 2013 年 4 月 1 日から 2014 年 6 月 20 日にニコニコ動画より収集した 1,758,322 件の動画と，182,135 件の動画リストを用いて行った．

入力キーワードとの関連度の評価は 8 人の被験者が各動画を再生，閲覧し主観により(A 評価)検索語と深い関係がある動画[+2 点]，(B 評価)検索語と関連があるが関係が深くない動画[+1 点]，(C 評価)検索語と無関係の動画[±0 点]の 3 段階の評価に分類した．評価者には，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(入力キーワード)は「ASKA」，「バルス」，「原爆」，「地震」，「27 時間テレビ」，「世界遺産」，「チャーハン」，「MTG(Magic: The Gathering)」，「政治家 A」とした．前半の 5 語は 2013 年 8 月 1 日から 8 日の検索エンジンにおける急上昇ワード[13]で 1 位の単語である．後半の 4 語は単語が多くの分野に分散する様に我々が主観で選択した単語である．この期間の検索エンジンにおける急上昇ワードは他に「松浦亜弥」，「炎の神秘龍」があったが前者は Web 検索エンジンの検索結果が 50 件に満たないため，後者はすべての被験者がこの単語に関する知識がなく主観評価を行うことができないため評価から除外した．

nDCG[14]を用いての各検索語の評価の結果を図 4 に示す．縦軸の値は 8 人の被験者の評価の平均である．nDCG は式(11)により算出されるが，本稿の評価では $\mathrm{rel_i}$ はすべて 2 として(すべての動画が検索語(入力キーワード)と関連がある状態を理想として)評価を行った．

$$\left.\begin{aligned} \mathrm{nDCG_p} &= \frac{\mathrm{DCG_p}}{\mathrm{IDCG_p}} \\ \mathrm{DCG_p} &= \sum_{\mathrm{i}=1}^{\mathrm{p}} \frac{2^{\mathrm{rel_i}} - 1}{\log_2 \mathrm{i}} \end{aligned}\right\} \tag{11}$$

図 4 の評価結果より，tiHITS 手法は「MTG」と「政治家 A」を除く検索語(入力キーワード)で Web 検索エンジンよりも高い関連度を実現しており，「ASKA」，「原爆」，「地震」，「世界遺産」，「チャーハン」の 5 語において最も高い関連度を実現していることが分かり，提案手法(tiHITS 手法)は関連度の高い動画の検索に有効であると考えられる．提案手法同士を比較すると，平均にて tiHITS 手法が優れており，全ての検索語(入力キーワード)にて tiHITS 手法は他の提案手法と同等(3 語にてわずかに劣っている)か，大きく勝る結果を示しており，tiHITS 手法が優れると考えられる．

tiHITS 手法では TF-IDF を用いることで検索語(入力キーワード)における動画と動画リストの強さを定量化したため，他の提案手法よりも良い結果となったと考えられる．

動画コミュニティ抽出手法は Center 動画集合と Fan 動画

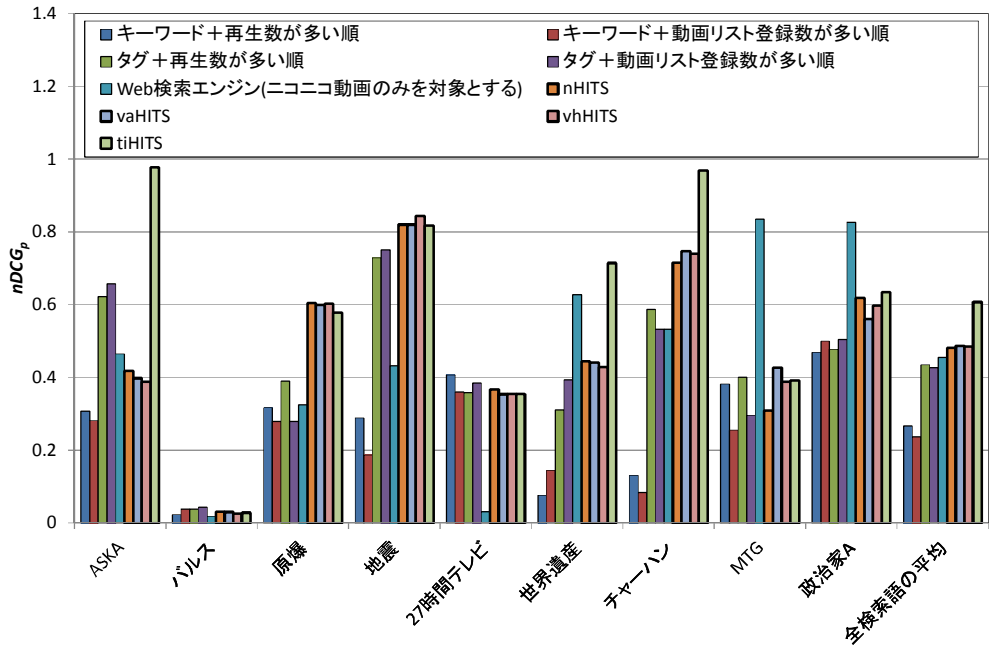

図 4 nDCG による評価結果

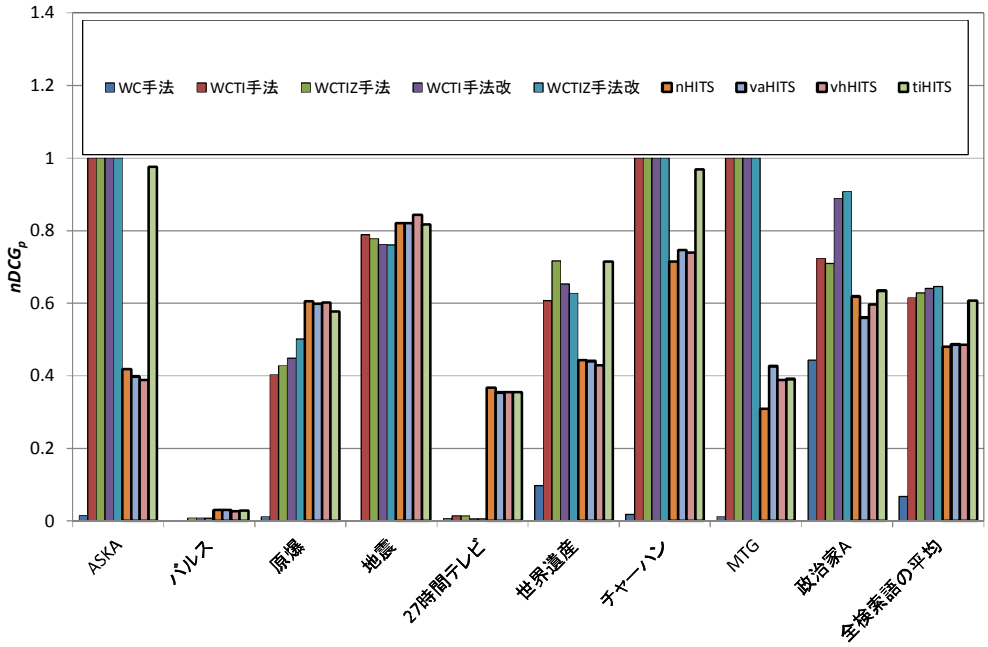

図 5 nDCG による評価結果(提案手法と動画コミュニティ抽出手法の比較)

リスト集合の更新を行っていく過程で新しい動画または動画リストが Center 動画集合または Fan 動画リスト集合に入るたびに集合を拡張していく手法であり，提案手法は baseset 集合を作成した後は baseset 集合内で反復計算を行い更新していくので baseset 集合からさらに拡張することはない．よって，動画コミュニティ抽出手法と提案手法は必ずしも直接比較できないが，参考のために両手法の検索結果の比較を図 5 に示す．

動画コミュニティ抽出手法は，抽出された動画コミュニティ内の動画の上位 50 件を検索結果としている．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ている．x(Center 動画集合と Fan 動画リスト集合の大きさ)は 100 件としている．図 5 より，動画コミュニティ抽出手法の方が提案手法よりも優れていることがわかった．これは，動画コミュニティ抽出手法は Center 動画集合と Fan 動画リスト集合の更新の際に集合が拡張されていくことでより大きな範囲で検索語と関係のある動画を探し出すことができるためであると考えられる．一方で，提案手法は baseset 集合からさらに拡張

することはないため，決められた範囲内でのみ検索語(入力キーワード)と関係のある動画を探すため，データがクロール済みでない状態でも比較的容易に適用できると考えられる．また，tiHITS 手法は動画コミュニティ手法に近い精度を提供できていることが分かる．これより，TF-IDF の動画リストへの適用が適合度の向上に大きな効果があると考えられる．

6. おわりに

本稿では，HITS アルゴリズムに着目し，これを応用した動画検索手法を提案した．評価の結果，HITS アルゴリズムと TF-IDF を用いた動画検索手法(tiHITS 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，有効性が確認された．

今後は，さらに多くの検索語による評価をし，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

本研究はJSPS 科研費 24300034, 25280022, 26730040 の助成を受けたものである．また，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] 西友規, 山口実靖, 小林亜樹, “動画リストの主題の偏りを利用した動画コミュニティ抽出”, 信学技報, vol. 113, no. 150, DE2013-28, pp. 163-168, 2013.

[4] 西友規，山口実靖，小林亜樹, "公開動画リストを用いた動画検索における Center 動画の選定に関する一考察", 情報処理学会 第 76 回全国大会, 5M-5.

[5] J. Kleinberg, “Authoritative Sources in a Hyperlinked Environment”, In Proceedings ACMSIAM Symposium on Discrete Algorithms, pp.668-677, 1988.

[6] P. K. Reddy, M. Kitsuregawa, “An approach to relate the web communities through bipartitegraphs,” Proc. of the 2nd International Conference on Web Information Systems Engineering, 2001.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] S. Baluja, R. Seth, D. Siva, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran and M. Aly, "Video Suggestion and Discovery for YouTube: Taking Random Walks through the View Graph", Proc. 895-904, 2008.

[13] Google トレンド - 急上昇ワード https://www.google.co.jp/trends/hottrends

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