6N-06

動画共有サイトにおける公開動画リストの主題の偏りに基づいた動画検索手法の選定

西　友規† 　山口　実靖† 　小林　亜樹†

†工学院大学大学院工学研究科電気・電子工学専攻

1. はじめに

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，動画共有サイトの動画の単語検索機能は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の向上は重要な課題の一つと考えることができる．

本稿では，既存の Web コミュニティ抽出手法と公開動画リストを用いた動画検索手法[2]において公開動画リストのタグ出現頻度を用いることで動画検索手法の選定についての考察を行う．

2. 動画コミュニティ抽出手法

2.1 WC 手法

WC 手法は，Web コミュニティ抽出手法[3]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(リンク先ページ)，Fan(リンク元ページ)，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以下の(1)～(4)に示す．

（1）共通の話題を持った動画を指定数(本稿の実験では 10 件)選択し，それを初期の Center 集合とする．

（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(本稿の実験では 100 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(本稿の実験では 100 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．

2.2 WCTI 手法

WCTI 手法[2]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，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．

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

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．

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

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．

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．

2.3 WCTIZ 手法

WCTIZ 手法[2]は，WCTI 手法に Zipf の法則を用いた手法である．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．

2.4 WCTI+手法 , WCTIZ+手法

WCTI+手法と WCTIZ+手法[2]では WCTI 手法と WCTIZ 手法における 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．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)を用いている．

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

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．

3. 動画検索手法の選定

前章で述べた動画コミュニティを用いる手法は，多くの検索語において動画共有サイトの検索機能より高い適合率で動画を検索できることが確認されているが，一部の検索語において動画共有サイトの検索機能より低い適合率となることが確認されている[2].

本章において，初期 Center 集合と抽出動画の評価を比較し，適切な動画検索手法を選定する手法を提案する．具体的には，初期 Center 動画から抽出された Fan 動画リストを式(1)を用いて評価し，次に式(2)とこの Fan 動画リストを用いて初期 Center 動画を評価する．そして，初期 Center 動画の評価値の合計と収束後の Center 動画の式(2)による評価値の合計を比較する．前者の評価値の方が高い場合は，動画共有サイトの検索機能による検索結果の方が適合率が高いと予想し動画共有サイトの検索機能に

Choosing Video Searching Method Based on Unity of Videos in Video Lists.

Yuki Nishi†, Saneyasu Yamaguchi†, Aki Kobayashi†

†Electrical Engineering and Electronics, Kogakuin University Graduate School

表1 初期 Center と収束後 Center の評価

検索語	初期Center	収束後Center
ASKA	10.00385491	100.6548362
バルス	10.00004931	6.000000001
原爆	10.00253382	23.00000001
地震	10.00013117	100.226481
27時間テレビ	10.38917143	9.000001844
世界遺産	10.00485613	51.00001293
チャーハン	10.00251442	54.0000803
MTG	10.34874226	100.0000351
政治家A	10.00325143	100.1894737

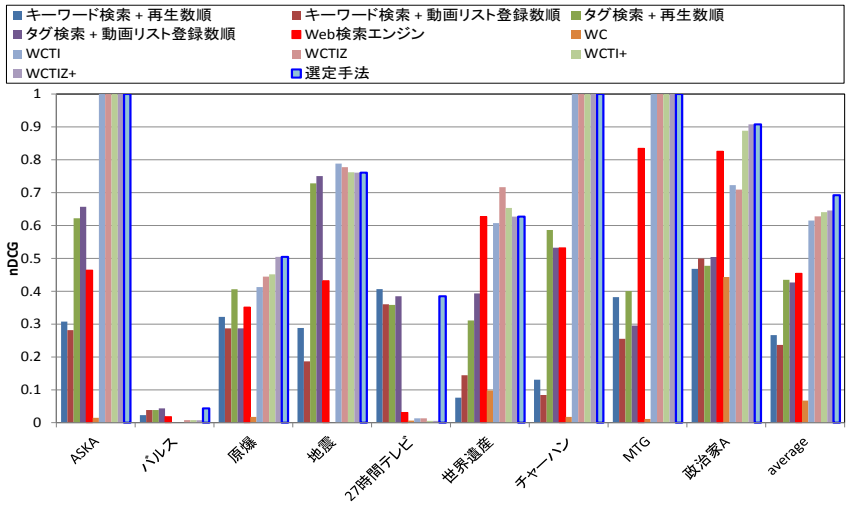

図 2 評価結果

よる検索結果を採用する．

4. 評価

本章では，動画共有サイトの検索機能，Web 検索エンジン，既存の動画コミュニティ抽出手法(WC 手法，WCTI 手法，WCTIZ 手法，WCTI+手法，WCTIZ+手法)，選定手法(WCTIZ+手法に 3 章の選定方法を適用)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．既存手法および選定手法では，抽出された動画コミュニティ内の動画の上位 50 件を検索結果とした．既存手法と選定手法の初期 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．また，選定手法において動画コミュニティを用いる手法の検索結果より動画共有サイトの検索機能の結果の方が優れていると判断された場合は，初期 Center 動画集合(タグ検索結果を動画リスト登録数順に並び替えたもの)を検索結果とする．動画共有サイトにはニコニコ動画を用い，抽出は 2013 年 7 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．

検索結果の評価は 4 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．

各手法における検索結果を nDCG[4]を用いて評価した．検索結果上位 p 件の nDCG の値(以下，$nDCG_p$ と記す)は以下の式(4)の様に算出される．

$$\left.\begin{aligned} nDCG_p &= \frac{DCG_p}{IDCG_p} \\ DCG_p &= \sum_{i=1}^{p} \frac{2^{rel_i}-1}{log_2 i} \end{aligned}\right\} \quad (4)$$

ただし，rel_iは各手法の順位 p における評価点数を表す．$IDCG_p$ は DCG_p が最大の値を取る場合を表すが，本稿では上位 p 件の検索結果がすべて検索語と深い関係がある動画を得られたものとしてrel_iの値を 2 に固定した．

表 1 に初期 Center 動画と収束後 Center 動画の評価値を示す．表より，検索語「バルス」と「27 時間テレビ」において初期 Center 動画からの劣化が確認できた．

図 2 に nDCG による評価結果を示す．各検索語の nDCG の値は 4 人の被験者の平均である．また，average は各検索語の平均の値である．図より，劣化が確認された検索語「バルス」と「27 時間テレビ」において動画コミュニティ抽出手法ではなく動画共有サイトの検索を用いた選定手法がより良い結果になっていることを確認できる．

5. おわりに

本稿では，既存の動画コミュニティ抽出手法において動画検索手法の選定を行った．評価の結果，選定を行うことにより既存手法よりもさらなる精度の向上が可能であることが確認できた．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．多くの有益な助言をくださりました東京大学の豊田正史先生に感謝いたします．

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] Yuki Nishi, Saneyasu Yamaguchi, Aki Kobayashi, "Video Search in Video Sharing Site based on Public Video Lists", 9th International Conference on Ubiquitous Information Management and Communication ACM IMCOM (ICUIMC) 2015.

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

[4] K. Jarvelin et al., Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, Vol.20, No.4, pages 422-446, 2002.