3N-5

動画リストを用いた動画共有サイトにおける動画検索

西　友規†　　山口　実靖†

†工学院大学　工学部　情報通信工学科

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，多くの動画が共有されている．しかし，動画共有サイトで提供されている検索機能の精度は必ずしも十分とは言えない．動画共有サイトにおける動画検索の質の向上が重要と考えられる．

本稿では動画共有サイトで提供されているタグと公開動画リストに着目し，既存の動画コミュニティ抽出手法[1]と TF-IDF を用いる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．

2. Web コミュニティ抽出手法

Web 空間の中には共通の話題を有する Web ページ群が存在し，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．Web 空間からの Web コミュニティ抽出に関しては多くの研究成果が得られている[2,3,4,5].

これらの手法では，特定の話題を持つ Web ページを 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，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．

3. 動画コミュニティ抽出手法

福井らは，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考え，既存の Web コミュニティ抽出手法を用いて動画コミュニティを抽出する手法を提案している[1]．本手法では，表 1 の対応により Web コミュニティ抽出手法を動画共有サイトに適用する．

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

公開動画リストを Fan，公開動画リストに登録されている動画を Center として，Center 動画集合から Fan 動画リストの抽出，Fan 動画リスト集合から Center 動画の抽出を繰り返し，動画コミュニティを抽出する．

A Video Search Method with Public Video Lists
Yuki Nishi†, Saneyasu Yamaguchi†
†Department of Information and Communications Engineering, Kogakuin University

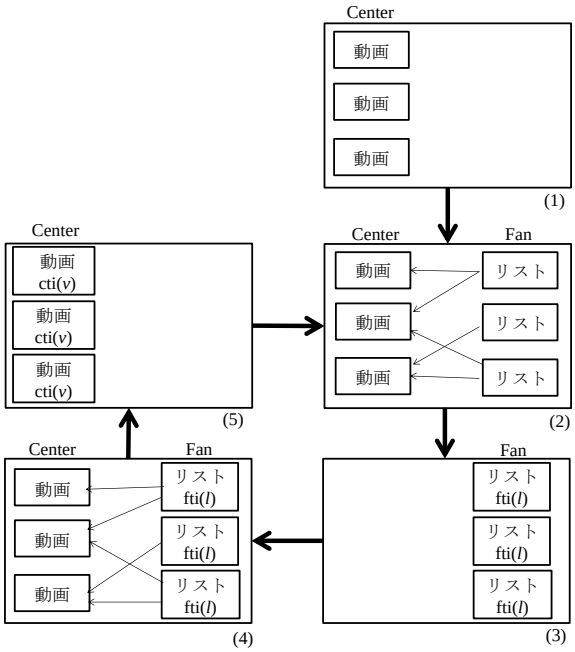

図 1 TF-IDF を用いる動画コミュニティ抽出法

4. 提案手法(WCTI 手法)

動画コミュニティ抽出手法と TF-IDF を用いた動画コミュニティ抽出手法を提案する．3 章の動画コミュニティ抽出手法同様に，本手法でも 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，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．

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

WCTI 手法における動画コミュニティ抽出手順を図 1 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 動画集合とする．

（2） Center 動画集合を登録している全動画リストを抽出し，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 100 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．10 乗とした理由は，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．

（3） Fan 動画リスト集合に登録されている全動画を抽出し，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 50 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．$\sum_{l \in L} \mathrm{fti}(l)$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．

（4） 収束をするまで，あるいは十分な回数，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，既存手法(動画コミュニティ抽出手法)，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．動画共有サイトの検索結果は，キーワード検索結果およびタグ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，の 4 通りを用いた．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 件を検索結果とした．また初期 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．抽出には，2012 年 7 月 1 日から 2012 年 12 月 10 日に収集した 1,453,300 件の動画と，124,523 件の動画リストを用いた．

検索結果の評価は，6 人の被験者（著者は含まれていない）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[+1 点]，(B 評価)検索語と関連があるが関係が深くない動画[±0 点]，(C 評価)検索語と無関係の動画[-1 点]，の 3 段階の評価に分類した．評価結果を表 3 から表 6 に示す．

表 3 は検索語「世界遺産」の検索結果を被験者らが評価を行った平均値である．同様に表 4 は「チャーハン」，表 5 は「MTG」，表 6 は「政治家 A」の検索結果を被験者らが評価を行った平均値である．

表 3 評価結果:検索語「世界遺産」

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	2.00	10.33	37.67	−35.67
キーワード検索＋動画リスト登録数が多い順	4.67	10.00	35.33	−30.67
タグ検索＋再生数が多い順	12.67	16.67	20.67	−8.00
タグ検索＋動画リスト登録数が多い順	15.00	14.67	20.33	−5.33
検索エンジン(動画共有サイトのみを対象とする)	27.00	11.00	12.00	15.00
既存手法(動画コミュニティ抽出手法)	2.00	12.33	35.67	−33.67
提案手法(WCTI手法)	40.67	7.00	2.33	38.33

表 4 評価結果:検索語「チャーハン」

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	5.50	2.00	42.50	−37.00
キーワード検索＋動画リスト登録数が多い順	2.75	2.25	45.00	−42.25
タグ検索＋再生数が多い順	21.75	5.50	22.75	−1.00
タグ検索＋動画リスト登録数が多い順	23.50	6.00	20.50	3.00
検索エンジン(動画共有サイトのみを対象とする)	23.75	2.25	24.00	−0.25
既存手法(動画コミュニティ抽出手法)	0.25	0.75	49.00	−48.75
提案手法(WCTI手法)	39.75	3.50	6.75	33.00

表 5 評価結果:検索語「MTG」

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	10.33	26.67	13.00	−2.67
キーワード検索＋動画リスト登録数が多い順	5.00	25.00	20.00	−15.00
タグ検索＋再生数が多い順	10.33	28.00	11.67	−1.33
タグ検索＋動画リスト登録数が多い順	5.67	28.33	16.00	−10.33
検索エンジン(動画共有サイトのみを対象とする)	38.33	10.00	1.67	36.67
既存手法(動画コミュニティ抽出手法)	0.00	1.33	48.67	−48.67
提案手法(WCTI手法)	45.67	1.33	3.00	42.67

表 6 評価結果:検索語「政治家 A」

	A(+1)	B(±0)	C(−1)	合計
キーワード検索＋再生数が多い順	17.00	21.67	11.33	5.67
キーワード検索＋動画リスト登録数が多い順	17.33	21.33	11.33	6.00
タグ検索＋再生数が多い順	17.67	21.67	10.67	7.00
タグ検索＋動画リスト登録数が多い順	18.00	21.67	10.33	7.67
検索エンジン(動画共有サイトのみを対象とする)	34.67	11.33	4.00	30.67
既存手法(動画コミュニティ抽出手法)	16.33	18.67	15.00	1.33
提案手法(WCTI手法)	50.00	0.00	0.00	50.00

すべての評価結果において，動画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，最も(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．

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(WCTI 手法)を提案した．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，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

本研究は科研費（22700039），（24300043）の助成を受けたものである

参考文献

[1] 福井紀彦，山口実靖，“動画共有サイトにおけるコミュニティ抽出”情報処理学会 2012 年全国大会 6N-7

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

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

[5] Gary William Flake, Steve Lawrence, C. 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.