D-006

動画リストと動画で構成される有向グラフを用いた動画検索に関する一考察

Movie Searching based on Directed Graph Composed of Movies and Movie 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，Web ページ間のリンク構造を解析してランキングを行う HITS アルゴリズム[2]に着目し，これを応用した動画検索手法を提案する．そして，評価実験の結果を示し提案手法の有効性を示す．

2．HITS

HITS アルゴリズム[2]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(1)，式(2)で求められる．また，式(3)と式(4)により正規化される．

$$x_i = \sum_{p_j \to p_i} y_j \quad (1)$$

$$y_i = \sum_{p_i \to p_j} x_j \quad (2)$$

$$x_i = \frac{x_i}{\sqrt{\sum_j^n x_j^2}} \quad (3)$$

$$y_i = \frac{y_i}{\sqrt{\sum_j^n y_j^2}} \quad (4)$$

3．提案手法

3.1. 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．具体的な手法は次章にて述べる．

（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(1)を用いて baseset 集合内の動画に authority の評価値x_iを求め，式(3)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(2)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(4)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．

3.2. 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．3.1 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(1)ではなく式(5)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．

vhHITS 手法では公開動画リストの評価を求める際に式(2)ではなく式(6)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．

$$x_i = \sum_{p_j \to p_i} y_j \times v_i \quad (5)$$

$$y_i = \sum_{p_i \to p_j} x_j \times v_j \quad (6)$$

†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

3.3. 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．3.1 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(1)ではなく式(7)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．

$$x_i = \sum_{p_j \to p_i} y_j \times tfidf_j \qquad (7)$$

4. 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，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．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．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．

検索結果の評価は 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．評価結果を表 1 に示す．検索語は「ASKA」，「バルス」，「原爆」，「地震」，「２７時間テレビ」，「世界遺産」，「チャーハン」，「MTG」，「政治家 A」とした．前半の 5 単語は 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[3]で 1 位の単語である．後半は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．表 1 は，すべての語の主観評価の平均である．

表 1 TF-IDF を動画共有サイトに適用

	A[+2]	B[+1]	C[±0]	合計
キーワード検索＋再生数が多い順	9.5	13.3	24.9	32.3
キーワード検索＋動画リスト登録数が多い順	8.2	12.1	27.4	28.6
タグ検索＋再生数が多い順	17.1	16.3	14.4	50.5
タグ検索＋動画リスト登録数が多い順	15.8	14.9	17.1	46.4
nHITS	20.3	12.9	14.5	53.6
vaHITS	20.6	12.7	14.5	53.8
vhHITS	20.2	13.7	13.8	54.2
tiHITS	25.1	13.5	9.1	63.8

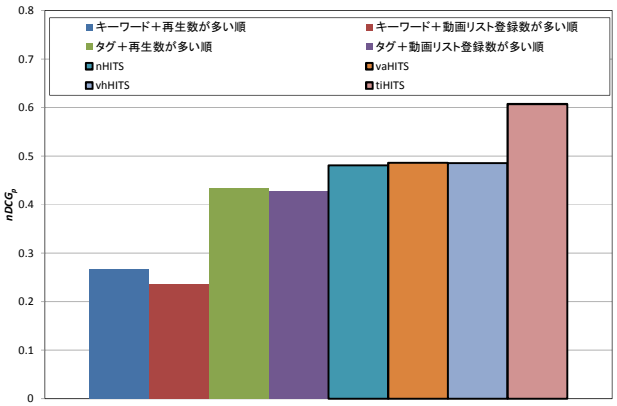

図1 nDCG による評価

また，nDCG[4]を用いての評価の結果を図 1 に示す．nDCG は下記の式(8)により算出されるが，本稿の評価では rel_i はすべて 2 として(すべての動画が検索語と関連がある状態を理想として)評価を行った．表 1，図 1 より，tiHITS が他の方法より優れており，検索語と関連の高い動画を抽出できることが確認された．

$$\left.\begin{aligned} \mathrm{nDCG_p} &= \frac{\mathrm{DCG_p}}{\mathrm{IDCG_p}} \\ \mathrm{DCG_p} &= \sum_{\mathrm{i=1}}^{\mathrm{p}} \frac{2^{\mathrm{rel_i}} - 1}{\log_2 \mathrm{i}} \end{aligned}\right\} \qquad (8)$$

5. 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，有効性を確認した．

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

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

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

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