

動画コンテンツ共有サイトの可視化手法の研究

上江まり子[†] 橋本 隆子^{††} 北川 博之^{†††}

[†] 筑波大学第三学群情報学類 〒 305-8573 茨城県つくば市天王台 1-1-1

^{†††} 筑波大学大学院システム情報工学研究科 〒 305-8573 茨城県つくば市天王台 1-1-1

^{††} 千葉商科大商経学部 〒 272-8512 千葉県市川市国府台 1-3-1

E-mail: [†]kamie@kde.cs.tsukuba.ac.jp, ^{††}takako@cuc.ac.jp, ^{†††}kitagawa@cs.tsukuba.ac.jp

あらまし 利用者自らが動画コンテンツを投稿でき、また投稿されたコンテンツを視聴してコメントを付加できる動画コンテンツ共有サイトは近年大きな広がりを見せている。利用者数は劇的に増加し、投稿されるコンテンツ数も増加の一途をたどっている。本稿ではこのような動画コンテンツ共有サイトに着目する。蓄積された膨大な量の動画コンテンツ集合を解析してトピックを抽出、トピックの特性分析を行い、動画コンテンツ集合の構造化を図る。構造化された情報をトピック単位で可視化し、利用者に動画コンテンツ集合全体の概要把握を促すようなインタフェースの提供を目指す。

キーワード 可視化、動画共有サイト、トピック抽出、クラスタリング

Visualization for Video-Sharing System

Mariko KAMIE[†], Takako HASHIMOTO^{††}, and Hiroyuki KITAGAWA^{†††}

[†] College of Information Sciences, University of Tsukuba Tennodai1-1-1, Tsukuba-shi, Ibaraki, 305-8573 Japan

^{†††} Graduate School of Systems and Information Engineering, University of Tsukuba Tennodai1-1-1, Tsukuba-shi, Ibaraki, 305-8573 Japan

^{††} Faculty of Commerce and Economics, Chiba University of Commerce Konodai1-3-1, Ichikawa-city, Chiba, 272-8512 Japan

E-mail: [†]kamie@kde.cs.tsukuba.ac.jp, ^{††}takako@cuc.ac.jp, ^{†††}kitagawa@cs.tsukuba.ac.jp

Abstract Video-sharing services, which allow users to upload video clips to an internet website, have rapidly spread recently. Through the services, users can view videos and add comments to them. The number of users in the services has increased dramatically as well as the number of videos. This paper focuses on such video-sharing services and proposes a visualization method for huge quantities of video clips on the services. Our proposed method extracts topics through analysis of video clips' meta-data and provides an interface to grasp the outline of the extracted topics. We also present our experimental results of applying our proposed methods for existing video sharing service.

Key words Visualization, Video-sharing system, Topic detection, Clustering

1. 序 論

動画コンテンツ共有サイト（以下、動画共有サイトと記す）とは、利用者自らが動画コンテンツ（以下、コンテンツと記す）を投稿でき、投稿されているコンテンツを視聴することができる Web サイトである。この動画共有サイトは、近年大きな広がりを見せている。利用者数が劇的に増加しており、YouTube [1] やニコニコ動画 [2] といった、代表的な動画共有サイトでは 2009 年 2 月時点でそれぞれ、2170 万人/月、1120 万人/月のユーザ

アクセスを達成している（日本国内）。このような動画共有サイトの普及を受け、企業や政治といった各方面からの注目が高まっている。複数の企業は既にプロモーションビデオや予告編を動画共有サイト上で配信し、宣伝戦略に活用している。政治の分野では民主党を筆頭に、主要政党のほとんどが“YouTube”や“ニコニコ動画”等に公式チャンネルを持っている。

このような状況の中、我々は、“増加し続ける膨大な量のコンテンツとその情報量を生かした、新しいコンテンツ視聴形態”、について考える。

我々は検索結果として多数のコンテンツが得られた場合、検索結果コンテンツ集合の中に複数のトピック（話題、話のまとまり）が存在すると考える。検索結果コンテンツ集合の中にどのようなトピックが存在し、各トピックがいつ・どの位注目されたのか、という情報は、利用者のコンテンツ選択に有益な情報と考える。しかし既存の動画共有サイトの検索結果提示方法である“コンテンツの一覧表示”では、検索結果コンテンツ集合の中に存在するトピックを把握することは難しい。

そこで本稿では動画共有サイトの提供する検索結果コンテンツ集合の、構造化・可視化を行うシステムを提案する。システムは

- 検索結果コンテンツ集合からのトピック抽出・トピック特性分析による、検索結果コンテンツ集合の構造化
- 構造化した情報の可視化

を行う。本システムによって利用者は検索結果全体の概要を把握でき、その上で希望するコンテンツを探索することが出来る。これによって、動画共有サイトに蓄積されている膨大な情報を生かすことが出来ると思う。

2. 前 提

2.1 動画コンテンツ共有サイト

動画共有サイトとは利用者が自由にコンテンツを投稿・閲覧・共有できる Web サイトである。投稿されたコンテンツは世界中の利用者からアクセスが可能となる。利用者が投稿する形式から“動画投稿サイト”、との呼称もあるが本稿では“動画コンテンツ共有サイト”、あるいは“動画共有サイト”と表記する。

2005 年辺りからストリーミング配信形式の動画共有サイトがインターネット上に増加し、2006 年に米国の YouTube が注目を浴びたことで動画共有サイトは一挙に一般的なサービスとして普及していった。2007 年 11 月現在、国内外で数十単位の動画共有サイトが運営されている。一般に動画共有サイトでは閲覧のみであれば会員登録は不要であり、コンテンツをアップロードする際に無料の会員登録を必要とする場合が多い。会員登録は容易で誰でも簡単に登録することが出来る。コンテンツを投稿しなくても、会員登録を行うと他の利用者のコンテンツにコメントを付けたり、お気に入りのコンテンツのリストが作成出来たりする。また動画共有サイトには、社会的ネットワークをインターネット上で構築する、ソーシャルネットワーキングサービスの側面を持っているものも多い。例えば、コンテンツを共有することで他の利用者とコミュニケーションをとれたり、家族や友人のみに公開設定を出来たりするサイトが存在する。

3. 問 題 提 起

我々は検索結果として多数のコンテンツが得られた場合、そのコンテンツ集合の中には複数のトピック（話題）が存在すると考えている（例えば、“選挙”というキーワードで検索を行った結果には、

- 東京都議会選挙（2009 年 7 月 12 日施行）
- 衆議院議員総選挙（2009 年 8 月 30 日施行）

• 複数の県で施行された県知事選挙
といった、トピックが存在すると考えられる。）検索結果コンテンツ集合の中に

- どのようなトピックが存在するのか
- 各トピックはいつどの位注目されたのか

という情報は、利用者のコンテンツ選択に有益な情報と考えられる。しかし、そのようなトピックの存在は、既存の動画共有サイトの検索結果提示方法である“コンテンツの一覧表示”では把握し難い（図 1）。たとえ何万件ものコンテンツが検索にヒットしても、多数のコンテンツを視聴するには膨大な時間が必要となるため、利用者が実際に利用するのは一覧で提示された複数コンテンツのみが中心になる。したがって検索に該当したコンテンツの一覧表示では、動画共有サイトの情報量が生かされていないと考えられる。

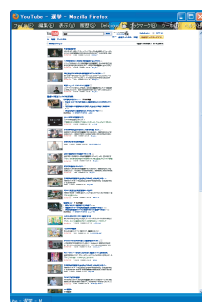


図 1 検索該当コンテンツの一覧表示

4. システム提案

我々は、3. 章の問題提起を受け、動画共有サイトの提供する検索結果コンテンツ集合の、構造化・可視化を行うシステムを提案する。システムは

- 検索結果コンテンツ集合からのトピック抽出・トピック特性分析による、検索結果コンテンツ集合の構造化
 - 構造化した情報の可視化
- を行う。

システムの構成は以下となる（図 2）。

- (1) コンテンツ収集モジュール
- (2) データ分析モジュール
- (3) 可視化モジュール

我々は図 2 のようなシステムを想定し、コンテンツ収集モジュールの開発、システムの鍵となるデータ分析モジュールと可視化モジュールの実現方法を検討した。以下、4.1～4.4 節でシステムが対象とする動画共有サイト、各モジュールの実現方法、実際に開発したモジュールの開発環境について述べる。

4.1 対象とする動画コンテンツ共有サイト

本研究が対象とするのは、以下の条件を満たす動画コンテンツ共有サイトである。

条件 1 コンテンツ毎に以下のメタデータを保有している。

- タイトル, サムネイル, 説明文, タグ情報, 投稿日, 再生回数, 投稿者名

条件 2 条件 1. に示したメタデータの取得が可能である。

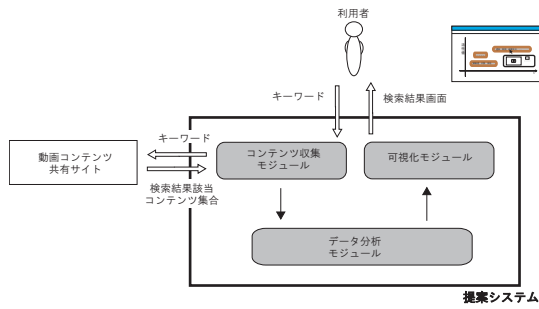


図2 提案システムの構成

本稿では YouTube を対象とし、研究を行った。

4.2 収集モジュール

コンテンツ収集モジュールでは、以下の動作を行う。

(1) 利用者が入力したキーワードから動画共有サイト上で検索を行う。

(2) 動画共有サイトが提供する、検索結果該当コンテンツ集合のメタデータを取得する。取得するメタデータは 4.1 の条件 1. に示した 7 つのデータとする。

我々は以下の環境で、本稿で対象とする動画共有サイトである “YouTube” 専用の、コンテンツ収集モジュールを開発した。

- (1) J2SE 5.0
- (2) JRE1.6.0_07
- (3) eclipse 3.2

4.3 分析モジュール

データ分析モジュールでは、コンテンツ集合の構造化を図る。具体的には、以下を行う。

1) コンテンツ集合からのトピック抽出

検索結果コンテンツ集合から、クラスタリングによってトピックを抽出する。クラスタ数はコンテンツ数に対する割合 (κ %) とした

2) トピックの特性分析

抽出した各トピックに対し、以下の特性を算出する

- 2-1) 活性値
- 2-2) トピック有効期間
- 2-3) トピックを代表するコンテンツ
- 2-4) トピックワード集合

次に、各分析の実現方法について順に述べる。

4.3.1 トピック抽出

本研究では、コンテンツの

- テキストデータ (タイトル・説明文・タグ情報)
- 時間データ (投稿日情報)

を利用し、コンテンツ集合からトピックを抽出できると仮定する。その理由を以下に示す。

【理由 1】コンテンツの投稿者は一般に、“コンテンツをより多くの人に見てもらいたい、気に入ってもらいたい”，という願望を持っている。そのため、コンテンツのタイトル・説明文・タグには、コンテンツの内容が容易に把握できるような、コンテンツを容易に見発できるような、的確な表現を心がける。

【理由 2】何らかのイベントが発生した時、そのイベントに

関連したコンテンツを投稿する人が急増する。従って、同時期に投稿されたコンテンツは同じトピックに関するものである可能性が高い。

我々は、理由 1, 理由 2, の考えに基づいたコンテンツ間の類似度をそれぞれ求める。理由 1, 理由 2 に対応する類似度を、それぞれ “テキストデータ類似度”, “時間的類似度” とする。その 2 種類の類似度を融合して求めたコンテンツ間の類似度を利用し、クラスタリングを行う。その結果形成された各クラスターを、一つのトピックと見なす。クラスタリングは階層型クラスタリングの Ward 法を使用した。

次に、“テキストデータ類似度” と “時間的類似度” の算出方法についてそれぞれ説明し、その後、2 つの類似度の融合方法を述べる。

• コンテンツ間のテキストデータ類似度

本類似度は、コンテンツのテキストデータ (タイトル・説明文・タグ情報) から算出する。我々は、テキストデータの表現する話題は、テキストに含まれる名詞により推測できると考える。従ってコンテンツのテキストデータから名詞を抽出し、抽出した名詞の出現頻度ベクトルを、コンテンツのベクトルと定義する。コンテンツ間の距離は、このコンテンツベクトルのコサイン類似度で求められる。

• コンテンツ間の時間類似度

理由 2 に基づき、コンテンツ間の投稿日の差から算出する。“コンテンツ間の投稿日が一定時間離れる度に、一定の割合でコンテンツ間の類似度が減少する” という仮定をおき、“忘却の概念を取り入れた時間類似度のモデル [12]” を導入する。コンテンツ間の投稿日の差が t の場合の時間類似度 $\text{TimeWeight}(t)$ を次式で定義する。

$$\text{TimeWeight}(t) = T_0 \times \exp\left(-\frac{0.693}{t_{1/2}}t\right) \quad (1)$$

本稿では、パラメタ T_0 と $T_{1/2}$ はそれぞれ 1 と 31 として実験を行った。従って、同じ日に投稿されたコンテンツの時間類似度は 1。コンテンツの投稿日が 1 か月 (31 日) ずれている場合のコンテンツ間の時間類似度は 0.5 となる。

• コンテンツ間の類似度

コンテンツ間の類似度は、テキストデータ類似度と時間類似度の重み付き平均とする。

$$\text{類似度} = \text{テキストデータ類似度} \times m + \text{時間類似度} \times n \quad (2)$$

4.3.2 トピックの活性値

トピックの活性値は、トピックが動画共有サイト上でどれだけ話題になったかを示す特性である。

トピック s の活性値 $\text{Activity}(s)$ を、トピック s に関連するコンテンツ集合 C_s の各コンテンツの活性値 $\text{activity}(s, c_j)$ の和とする。

$$\text{Activity}(s) = \sum_{c_j \in C_s} \text{activity}(s, c_j) \quad (3)$$

各コンテンツの活性値を求める際、本研究では次の 2 点を考慮

に入れた．

(1) 再生回数の多い動画の活性値は高い．

(2) あるトピックに関連するコンテンツ集合について，一人の投稿者が大多数のコンテンツを投稿している場合よりも，投稿者がばらついている場合の方が，より多くの人がそのトピックに関心があると見なせるので活性値が高い．

上記の(2)は，同一投稿者による類似コンテンツの連続投稿や，規定外の長さのコンテンツを複数のコンテンツに分割して投稿するような“シリーズもの”がトピックの活性値を不適切に高くする現象を緩和する効果もある．2を考慮するため， $IAF(Inverse\ Author\ Frequency)$ というパラメータを以下のように定義する．

$$IAF(s, c_j) = \frac{1}{\log\left(\frac{n(s, c_j)+1}{N(s)} + 1\right)} \quad (4)$$

$n(s, c_j)$ はコンテンツ c_j の投稿者がクラス s に投稿しているコンテンツの数， $N(s)$ はクラス s に関連するコンテンツの総数である．

この2点を踏まえ，トピック s のコンテンツ $c_j (c_j \in C_i)$ の活性値 $activity(s, c_j)$ を次式で定義する．

$$activity(s, c_j) = \log(view(c_j) + 1) \times IAF(s, c_j) \quad (5)$$

4.3.3 トピック有効期間

トピックが話題となっている期間をトピック有効期間とする．トピックに関連するコンテンツが投稿されている期間（最も昔に投稿されたコンテンツの投稿日から，最も最近投稿されたコンテンツの投稿日まで）とする．

4.3.4 トピックを代表するコンテンツ

クラスタのセントロイドとその近隣4コンテンツ，合わせて5コンテンツをトピックを代表するコンテンツとする．

4.3.5 トピックワード集合

トピックワード集合とは，トピックを代表する単語集合を示す．このトピックワード集合を，トピックを代表するコンテンツ集合内で，全単語について $TF-IDF$ 値の和を求め，合計値が高い上位10単語と定義する．

4.4 可視化モジュール

我々は，検索結果コンテンツ集合の解析結果を以下のように可視化する

- 検索キーワードに関連するトピックを提示．
 - － トピックを活性値-時間グラフ上に示す．利用者はトピックが出現し話題になった時期とトピックの活性値を一目で把握出来る．
- トピックのトピックワード集合を提示．
 - － トピックについて，そのトピックを最もよく表現するコンテンツ（以後，トピックを代表するコンテンツ，または代表コンテンツと記す）を提示可能．代表コンテンツのタイトル，説明文，投稿日，リンク，サムネイルを表示．

例えば，“小沢一郎”というキーワードで検索を行った結果，図3のような画面を提示する．図3では，

- 衆議院議員総選挙
- 東京都議会議員選挙

という2つのトピックがそれぞれ，“選挙，衆院，政権交代”，“都議選，代表，辞任”というトピックワード集合で表現されている．各トピックのグラフ座標から，

- “衆議院議員総選挙”が“東京都議会議員選挙”の後に施行されたこと
- “衆議院議員総選挙”が“東京都議会議員選挙”に比べて，より動画共有サイトで話題になったことが即座に把握できる．

このような可視化により，利用者は検索結果コンテンツ集合の傾向，トピックを容易に把握出来る．その上で，興味のあるトピックを選択し，コンテンツを探索していくことが出来る．

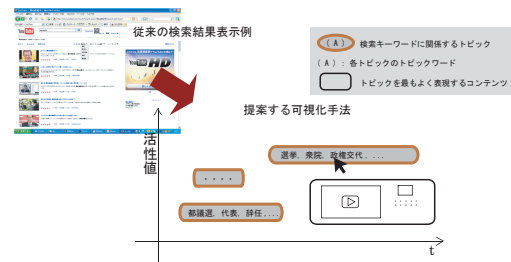


図3 提案可視化手法

我々は実際に@nifty TimeLine を利用し，検索結果の可視化を行った（図5）．グラフの縦軸・横軸は図3と同様になっている．グラフ上に，示された線分がトピックに対応している．トピック毎に吹き出しでトピックワード集合を提示した．トピック間の時間関係の把握を容易にするため，トピック間を時間系列で結んだ．

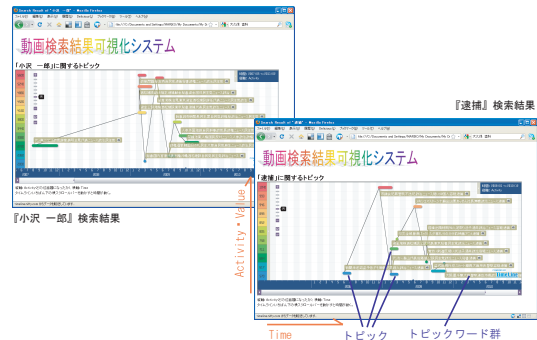


図4 @nifty TimeLine を利用して実現した提案可視化 —全体図—

トピックにマウスカーソルを合わせると，トピックを代表するコンテンツ集合のサムネイルが表示される．トピックをクリックすると，トピックの詳細として以下が表示される（図5）．

- トピックワード集合
- 有効期間
- 代表コンテンツ集合のタイトル
- 代表コンテンツ集合のサムネイル
- 代表コンテンツ集合のリンク

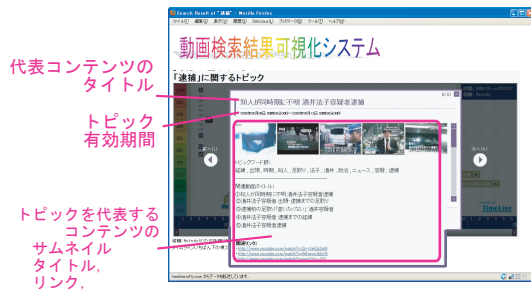


図 5 @nifty TimeLine を利用して実現した提案可視化 —トピック選択時—

5. 評価実験

提案した検索結果可視化方法の有用性を測るために評価実験を行った。以下の検索キーワードについて実際にトピック抽出・特性分析を行い、検索結果をトピック単位で可視化した。

キーワード：

逮捕 (1003), 殺人 (1009), 不況 (1095), 小沢 一郎 (1147)
鳩山 由紀夫 (1161), 浅田 真央 (950), オリンピック (1468)

キーワード毎についている数字は、YouTube から取得したコンテンツ数のデータである（外国語コンテンツを除く）。トピック抽出の際に必要なパラメータは次のように設定した。

- $\kappa=7.5$
- $m=0.7, n=0.3$

各トピックに対して求めた“活性値”、“トピックワード集合”、“トピックを代表するコンテンツ”のうち、検索キーワード“逮捕”に関して、活性値の高かった上位 5 トピックを示す（表 1）。活性値が 2 番目に高いトピックは外国人犯罪に関するトピック、3 番目は“逮捕しちゃうぞ”というアニメに関するトピック、4 番目が酒井法子逮捕に関するトピック、5 番目が小沢一郎の公設第一秘書逮捕に関するトピックであることが推測できる。

分析した結果を可視化し、被験者 8 名による評価実験を行った。評価指標は以下とする。

- トピック把握：検索結果に、どのようなトピックが存在するかを把握できるか
- トピックの特性把握：トピックがいつどの位活発に取り上げられたか、把握できるか
- トピックの代表コンテンツ：トピックにどのようなコンテンツが含まれているか、把握できるか

評価値は 1~5 の 5 段階とし、3 を“どちらかと言えば可”、可視化がより妥当であると感じられたならば、より高い評価値を選択してもらうこととした。既存可視化方法（検索該当コンテンツの一覧表示）を比較対象とし、同様の評価を行った。

評価結果を表 2 に示す。表 2 の値は、被験者 8 名の評価結果の平均値である。既存動画共有サイトでの可視化手法を“既

表 1 “逮捕”に関する、活性値の高い上位 3 トピック

*	活性値, トピック発生時期	
	トピックワード トピックを代表するコンテンツのタイトル	
1	1035, 2008/12/19 ~ 2009/2/13	市議会, 犯罪, 警察, 不法, 犯, 政治, ニュース, 疑い, 中国人, 容疑, 逮捕
		不法就労させた中国人経営者の女ら 2 人逮捕
2	935, 2009/6/14 ~ 2009/7/20	メキシコ, マスター, シナ, 輸出, 凶悪, あっせん, 社長, 無職, 政治, ニュース, 逮捕
		逮捕の社長過去にミサイル関連機械輸出
3	763, 2009/3/29 ~ 2009/3/29	作文, 全域, 曾, 巣, ネックレス, 夕暮れ, 分の, 分け前, 映画, アニメ, 逮捕
		Taiho Shichauzo Episode 7 Part 1/3-1
4	737, 2009/8/6 ~ 2009/8/13	経緯, 出頭, 時期, 知人, 足取り, 法子, 酒井, 政治, ニュース, 容疑, 逮捕
		知人が同時期に不明 酒井法子容疑者逮捕
5	712, 2009/2/26 ~ 2009/3/25	資金, 地検, 西松建設, 公設, 代表, 東京, 秘書, 民主党, 政治, ニュース, 逮捕
		【国内】小沢代表 公設第一秘書逮捕 3 / 3 20090304

存”, 提案可視化手法を“提案”と記す。

表 2 検索結果可視化手法の評価実験

	トピック把握	特性把握	代表コンテンツ
提案	3.88	4.21	3.43
既存	3.37	2.62	3.25

全ての指標において、提案手法が既存の一覧表示を上回る評価を得た。また、被験者のコメントとして

- 話の固まりが理解しやすい
- 文字が少なく、見やすい
- 時系列表示が理解しやすい
- 新しい検索結果提示方法とを感じる

といった、提案可視化手法に好意的な意見が多く得られた。これらのことから、我々の提案するトピック単位の検索結果提示方法が有益であると考えられる。

6. 関連研究

動画共有サイトに注目が集まると共に、動画共有サイトを対象とした研究は盛んに行われており、

- 動画共有サイトで形成されているコミュニティの研究
- 動画共有サービスのシステム設計の研究
- 動画共有サイトのインターフェースの研究

など様々な種類の研究が存在する。

動画共有サイトのインタフェースに関する研究でも、

- 動画共有サイトの活性化を促すインタフェースの研究
- 動画共有サイトからの新たな情報取得を促すインタフェースの研究

など、数種類に分けることができる。

その中で、いくつかの関連研究を以下に挙げる。

- 動画共有サイトからの情報取得

動画共有サイトからの新たな情報取得を試みる研究として、平山ら [15] の研究がある。平山らは動画共有サイトにおいて、利用者が新たな発見を行う機会を提示するインタフェースを提案している。平山らが提案するのは、“受動的に情報を取得する”，動画共有サイトのインタフェースである。“受動的に情報を取得する”とは，“利用者からの操作を受けずにコンテンツが自動で次々と提示される”，ということである。つまり、テレビのつけっぱなし状態のような、システムが一方向的にコンテンツを提供し続けるような状態である。平山らは我々と同様に、動画共有サイトに蓄積されているコンテンツ情報を、新たなアプローチによって取得することを目的としている。しかし平山らは、個々のコンテンツ単位でのコンテンツ提示を行うが、我々はコンテンツ集合に含まれるトピックに着目し、トピック単位でのコンテンツ提示を行う。その点で平山らと我々の研究は着眼点が異なる。

- 検索結果 可視化

検索結果の提示方法としては、結城ら [16] の研究がある。結城らは Web 情報探索の支援のため、“検索キーワードに対する Web ページの概観の把握を促す”インタフェースを提案している。“検索キーワードに対する Web ページの概観の把握を促す”ため、結城らは検索キーワードの関連単語のボリューム、関連単語間の関係性を示している。“関連単語のボリューム”とは、検索キーワードと関連単語を合わせてキーワード検索にかけた（以下、二次検索と記す）結果得られるコンテンツ数を指す。結城らは検索キーワードの関連単語に焦点を当てて検索結果の可視化を行っている。これは我々が検索キーワードのサブトピックに焦点を当てて検索結果の可視化を行っていることと、同等と見なせる。また、結城らの考える“関連単語のボリューム”は、我々の“サブトピックの活性値”に相当すると考えられる。しかしながら、結城らが“関連単語のボリューム”を単純に二次検索の該当コンテンツ数から求めているのに対し、我々は“サブトピックの活性値”を、サブトピックに関連するコンテンツ数だけでなく、コンテンツの“再生回数”や“投稿者情報”といった、動画コンテンツ特有のパラメータを用いて算出している点で、両研究は異なる。

- 対象からのトピック抽出と可視化

平田ら [17] は可視化対象を、動画コンテンツではなくニュース記事とし、対象のトピック単位による可視化を行っている。平田らの研究は我々とトピックの抽出方法が同じであり、トピックに対して注目度（我々の求める“活性値”と同等と考えられる）、関連単語を算出している点で我々の研究と類似している。しかし、結城らの研究との違いと同様に、我々は動画コンテンツ特有のパラメータを用いて“活性値”を求める点が異なる。また、トピックの関連単語（我々の研究での“トピックワード”と同等と考えられる）の算出方法が異なる。結城らは、トピック内に出現する全名詞について、トピックに関連する記事の中で tf-idf 値の和をとり、その値の高かった上位数単語を関連単

語としている。しかし我々は各トピックのセントロイドに最も近い数コンテンツの間で、出現名詞の tf-idf 値の和をとり、トピックワードを決定している。

7. ま と め

動画共有サイトにおける、新たな検索結果提示手法を実現するシステムの提案を行った。コンテンツに付随する文字、数値情報を利用し、クラスタリングによりトピックを抽出した。更にトピック毎に特性分析を行い検索結果を構造化、検索結果をトピック単位で可視化した。評価実験を行い、可視化方法が我々の目的を達成していることを検証し、可視化手法に好感が持てるというコメントを多数得た。

今後は

- クラスタ数の再検討
- 固有名詞に重みづけ
- クラスタの凝集性の活性値への反映
- サムネイル画像の特徴の利用

を行い、トピック抽出・トピック特性分析の精度向上を行っていく。また被験者実験の結果から頂いたコメントを踏まえ、可視化手法の改善を試みる予定である。

謝 辞

本研究の一部は科学研究費補助金特定領域研究（21013004）による。

文 献

- [1] YouTube, YouTube, LLC, <http://www.youtube.com/>
- [2] ニコニコ動画 (9), ニワンゴ, <http://www.nicovideo.jp/>
- [3] 財団法人インターネット協会, インターネット白書 2009, インプレス R& D, 2009.
- [4] goo Research, NTT レゾナント, <http://research.goo.ne.jp/>
- [5] AmebaVision, 株式会社サイバーエージェント, <http://vision.ameba.jp/index.do>
- [6] フォト蔵, ウノウ株式会社, <http://photozou.jp/>
- [7] にゅーあきば. こむ, 株式会社にゅーあきば, <http://www.new-akiba.com/>
- [8] ワッチミー!TV, フジテレビラボ LLC 合同会社, <http://www.watchme.tv/>
- [9] ClipLife, 日本電信電話株式会社, <http://cliplife.jp/>
- [10] FlipClip, 株式会社フリップ・クリップ, <http://www.flipclip.net/>
- [11] Wikipedia フリー百科事典, ウィキメディア財団, <http://ja.wikipedia.org/wiki/>
- [12] 石川佳治, 北川博之, “忘却の概念に基づくクラスタリング手法の改良方式”, 日本データベース学会 Letters Vol. 2, No. 3, 2003.
- [13] ウェブテキスト解析システム『SlothLib』, <http://www.dl.kuis.kyoto-u.ac.jp/~ohshima/wiki/index.php?SlothLib>
- [14] @nifty TimeLine β, <http://timeline.nifty.com/>
- [15] 平山慧, 寺田実, 丸山一貴, “GutaGuta: 受動的な動画視聴と発見のためのインタフェース”, WISS2008
- [16] 結城崇, 三末和男, 志築文太郎, 高橋伸, 田中二郎, “Web 上のキーワードを辿る視覚的な情報探索インタフェースの開発”, 筑波大学第三学群情報学類平成 20 年度卒業論文
- [17] 平田紀史, 伊藤大樹, 大園忠親, 新谷虎松, “時系列を考慮した階層的クラスタリングに基づくインタラクティブなニュース記事閲覧支援システム”, The 23rd Annual Conference of the Japanese Society for Artificial Intelligence, 2009