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　†† 千葉商科大学 商経学部
††† 筑波大学大学院 システム情報工学研究科

1 序論

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．YouTube[1]やニコニコ動画[2]といった代表的な動画コンテンツ共有サイトは，それぞれ2170万人/月，1120万人/月のユーザアクセスを達成し（2009年2月時点），利用者は劇的に増加している．

現在の動画コンテンツ共有サイトの一般的な利用形態には，次のような場合分けが考えられる．

1. 見たいコンテンツが定まっている場合
2. 見たいコンテンツが定まっておらず，漠然とコンテンツを探したい場合

1の場合，検索キーワードを明示的に指定することで希望のコンテンツにたどり着ける．例えば，“小沢一郎が中国を訪問したことを報道するニュースコンテンツ”を検索するには，“小沢一郎 中国 訪問”という検索キーワードで絞り込めば良い．

一方2の場合には，検索キーワードを詳細に設定できないため，結果として多数のコンテンツが得られることとなる．例えば，“小沢一郎に関して話題になったトピックや各トピックの関連コンテンツを知りたい”という際に，“小沢一郎”とキーワード検索を行うと，YouTubeならば約1000件のコンテンツが得られる．利用者はそこから各動画投稿サイトが提供する“検索結果の並べ替え機能”を利用して希望のコンテンツを探索することになるが，実際に把握できるのは並び替えによって上位にランク付けされた少数の動画が中心となってしまう．これは希望する結果(“小沢一郎に関して話題になったトピックや各トピックの関連コンテンツの把握”)が得られていると言い難い上に，動画コンテンツ共有サイトに蓄積されている膨大な情報を生かし切れていない状況であると考えられる．

そこで我々は本稿で，多数のコンテンツが検索結果として得られた際に，それらを整理し，全体の概要や関連するトピックの推移などを把握できるような新しい視聴形態を提案する．

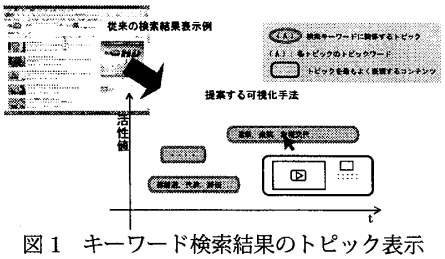

図1　キーワード検索結果のトピック表示

2 動画コンテンツ共有サイトの可視化

動画コンテンツ共有サイトの可視化手法として，以下を提案する（図1）

- 検索結果として得られたコンテンツ群から，検索キーワードに関連するトピックを抽出．
- 各トピックをトピックを象徴する単語（以下，トピックワード）群で表現．
- 各トピックを時間-活性値グラフ上にプロットし，動画コンテンツ共有サイト上でいつ頃，どれだけ注目されたかを提示．
- トピック毎に，そのトピックを最もよく表現するコンテンツ（以下，代表コンテンツ）を提示．

例えば，“小沢一郎”というキーワードで検索を行った場合，“衆議院議員総選挙”や“東京都議会議員選挙”という事柄がトピックとして抽出され，それぞれ“選挙，衆院，政権交代”，“都議選，代表，辞任”といったトピックワードで内容が説明される．更に時間-活性値グラフから，“東京都議会議員選挙”の後に“衆議院議員総選挙”が話題になっていること，“衆議院議員総選挙”のトピックの方がより話題になっていることなどが容易に把握出来るような可視化である．

3 動画コンテンツ共有サイト可視化システム

我々が開発した動画コンテンツ共有サイト可視化システムは，以下のモジュールから構成される．

1. コンテンツのトピック抽出
2. トピックワード群の抽出
3. トピックの活性値の算出
4. トピックを最も良く表現するコンテンツの選出

次に各モジュールについて説明する．

3.1 コンテンツのトピック抽出

本研究では，コンテンツのトピックはコンテンツのメタデータ（タイトル・説明文・タグ情報・投稿日情

A Visualization Method for Video-sharing Websites

Mariko KAMIE†(kamie@kde.cs.tsukuba.ac.jp)
Takako HASHIMOTO††(takako@cuc.ac.jp)
Hiroyuki KITAGAWA†††(kitagawa@cs.tsukuba.ac.jp)
†College of Information Sciences, University of Tsukuba, 305-8571, Ibaraki, Japan
††Commerce and Economics, Chiba University of Commerce, 272-8512, Chiba, Japan
†††Graduate School of Systems and Information Engineering, University of Tsukuba, 305-8573, Ibaraki, Japan

報）から抽出できると仮定する．なぜならば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うため，タイトルや説明文，タグにはコンテンツを的確に表す表現を入れる，”と考えられるからである．また，“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”とも考えられるため，投稿日情報も考慮に入れる．

我々のトピック抽出の手法を以下に示す．

1. コンテンツのテキストデータ（タイトル・説明文・タグ情報）から抽出した名詞群のtf-idf値を要素とする コンテンツベクトルを作成．コンテンツベクトル間の距離を計算し コンテンツ群の“テキストデータの距離行列”を算出．
2. コンテンツの投稿日の差からコンテンツ間の時間的距離を求め，コンテンツ群の“時間的距離行列”を作成[3]．
3. 1,2で作成した，各距離行列を融合してコンテンツ群の距離行列とし，これをクラスタリング．形成した各クラスタを一つのトピックと定義．

3.2 トピックワード群の抽出

各トピックのトピックワード群の抽出方法を示す．

1. “トピックに属するコンテンツ群のコンテンツベクトル”の重心を計算．
2. 重心と近接する5コンテンツを抽出．
3. 抽出した5コンテンツ内で，出現した各名詞のtf-idf値の和を計算．求めたtf-idf値の和が高い上位10単語をトピックワードと定義．

3.3 トピックの活性値の算出

トピックの活性値は，トピックに属するコンテンツの活性値の和とする．各コンテンツの活性値を求める際，本研究では次の2点を考慮する．

- 再生回数の多いコンテンツの活性値は高い．
- あるトピックに属するコンテンツ群において，少数の投稿者が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．

3.4 トピックを最も良く表現するコンテンツの選出

トピックに関連するコンテンツ群の中で，最も活性値が高いコンテンツとする．

4 実験

実際に動画コンテンツ共有サイトの検索結果に対してトピック分析を行い，提案手法で可視化する実験を行った．動画コンテンツ共有サイトの可視化は“@nifty TimeLine[4]”が提供するTimeLine上に時間-活性値グラフを作成することで行った．解析対象データは2009/12/13にYouTube上で“小沢 一郎”というキーワードで検索を行った結果得られたコンテンツ群935件である．抽出したトピックの内 活性値の高い上位10トピックを可視化した結果を図2に,表1に抽出したトピックの中で活性値が高い上位3件のトピックワード群，活性値，トピックが話題になった期間を示した．

図2で，各トピックはグラフ上に一本の線として表わされている．グラフの横軸と縦軸はそれぞれ時間と活性値を表わしており，各トピックが話題になった時期，話題になった度合いが瞬時に見てとれる．この可視化により，動画コンテンツ共有サイトが提供する“小沢 一郎“に関する大量のコンテンツ群の概要を利用者に提示出来ていると考えられる．

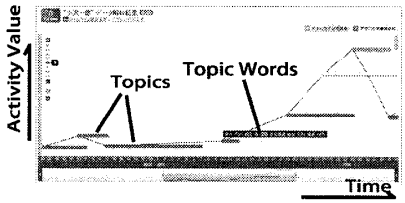

図2 データ可視化結果 (全体図)

表1 トピック抽出結果上位3件

#	トピックワード群	活性値	期間
1	党, 岡山, 加藤, 幸福, つかさ, 勝仁, 実現, 久我, 神奈川, 勝信	48.54	2009/8/13 ～ 2009/9/7
2	麻生, 交代, 卒業, 選, 出口, 自民党, 足寄, 衆院, 政権, 調査	44.81	2009/7/29 ～ 2009/8/18
3	管理, 保坂, 赤字, 法人, マスコミ, 厚, 外国, 運用, 年金, 積立	41.85	2009/7/18 ～ 2009/8/5

表1のトピック# 1,# 2は選挙に関するトピックであることがトピックワード群から考えられる．更にトピックの期間も考慮すると，# 1は第45回衆議院議員総選挙(2009/8/30 執行)，# 2は2009年東京都議会議員選挙(2009/7/21 施行)に関連するトピックであることが推測できる．またトピック# 3は年金積立金の赤字問題のトピックを示していると考えられるが，いずれも現状の結果ではトピックワード群が各トピックの内容を的確に表現しているとは言い難い．トピック及びトピックワードの抽出については，さらなる改良が必要であると考える．

5 まとめ

動画コンテンツ共有サイトに対して，蓄積された動画コンテンツを整理して利用者にわかりやすく提示することを目的とした動画コンテンツ共有サイト可視化手法の提案を行った．今後はトピック及びトピックワード群抽出の精度向上について検討を行うとともに，可視化手法について評価実験を行い，提案手法を改良していく予定である．

参考文献

[1] YouTube, http://www.youtube.com/
[2] ニコニコ動画(9), http://www.nicovideo.jp/
[3] 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.
[4] @nifty TimeLine β, http://timeline.nifty.com/