動画ニュース選択のためのニュース評価指標の構築に関する研究

Development of Evaluation Index for Adopting News Movies

中村 浩之†　小川 祐樹†　諏訪 博彦†　太田 敏澄†
Hiroyuki Nakamura　Yuki Ogawa　Hirohiko Suwa　Toshizumi Ohta

1. はじめに

世の中で起こった出来事を知る事のできる媒体として動画ニュースがある．日々起こる出来事は時々刻々とその情報を更新し続けているので，社会で起こっている出来事を把握したいという要求があっても，ニューストピックをチェックするために更新し続ける動画ニュースを全て見る事は膨大な時間が掛かる．

そこで，本研究は，社会で起こったニュース情報の移り変わりを把握し，視聴者が見たい動画ニュースを選択するための評価指標の構築を行う．同じシーンを集約するためにトピックを抽出し，トピックとシーンの特性を抽出する事で評価指標を構築する．動画ニュースに付随するテキストデータから話題となるトピック及びニュースシーンの抽出を行い，それらの内容を評価するために，共起タグを用いた時間変化するトピックの抽出と 4 つの評価指標によるニューストピックの特性と及びニュースシーンの特性を分析する事で評価指標の構築を目指す．

視聴者は大量の動画ニュースシーンから知りたい情報を自分で探し出す必要もなく，注目されたニュースや新しい情報を含むニュース、既知情報からの差分となるニュースの視聴が可能となる．

2. トピック抽出とニュース特性指標の提案

本研究では，動画ニュースをシーン単位で扱う．これまでの動画推薦システムは，番組単位で動画を推薦している事がほとんどで，そのため一つの推薦情報のなかに複数のニュース情報が含まれていた．図1で示すように，ニュース情報を類似する内容によって，点線枠で囲んだシーン集合のように，個別で扱う事によって，類似ニュースの放送頻度が把握できトピックの流行性が抽出できる．また，連続する日付で類似度が高いトピック同士を同じ話題を持つトピックとして処理をする事で新規性が抽出でき，トピックから重要なタグを抽出する事で要約性と進展性を抽出する事が出来る．

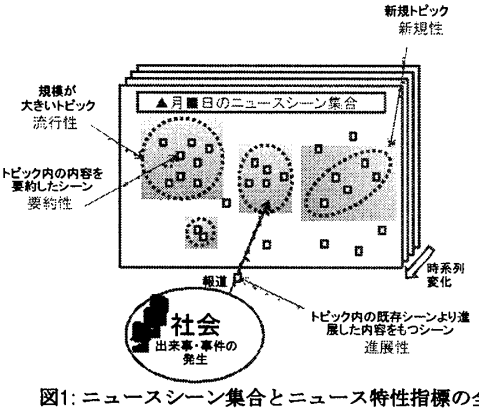

図1: ニュースシーン集合とニュース特性指標の全体像

以上4つの特性を分析するために，動画ニュースシーンに付随するテキストデータを以下の通りに処理を行う．

2.1. トピックの抽出

既にシーン単位で切り分けされた動画ニュースメタデータを対象に以下の処理を行う．

2.1.1. 用語の抽出

動画ニュースシーンに付随するヘッドラインなどのテキスト情報に対して，Mecabを用いた形態素解析を行い，テキスト内の名詞を対象とした単語及び複合語の出現頻度を計算する．

2.1.2. 特徴量計算

文章中の特徴的な名詞を抽出するため，tfidfを用いて，用語の特徴量の計算を行う．

2.1.3. 類似度ネットワーク分析

シーンごとにもっている用語の特徴量からシーン同士の類似性をcos類似度を用いて計算し，シーンtoシーンの類似度グラフを作成する．尚，類似度の閾値としてsim_{th}=0.3を暫定的に設ける．

$$sim(s_a, s_b) = \frac{\sum S_{i,a} S_{j,b}}{\sqrt{\sum (S_{i,a})^2} \sqrt{\sum (S_{i,b})^2}} \quad (1)$$

s_x：シーンx
$S_{x,y}$：シーンxにおけるタグyの特徴量

2.1.4. クラスタリングによるトピック抽出

Newman法を用いてシーン toシーン類似度ネットワークのmodularityの値が極大値になるエッジについて切り分けを行う事でクラスタリングをし，切り分けられたクラスタをトピックとして抽出する．

$$Q = \sum_{s=1}^{N_m} (e_{ij} - a_i^2) \quad (2)$$

N_m:ネットワーク内の全クラスタ数
e_{ij}:(クラスタiとクラスタjを結ぶエッジ数)÷全エッジ数
a_i:(クラスタiの持つ全てのエッジ数)÷全エッジ数

2.1.5. 時系列変化するトピックの定義

連続する日付で類似性が高いトピックを同じ話題をもつトピック群として定義する．トピックの類似性は3．1．3．と同様にcos類似度を用いて計算し，トピックtoトピックの類似度グラフを作成する事で同じ話題をもつトピッククラスタを抽出する．尚，類似度の閾値としてsim_{th}=0.6を暫定的に設ける．

$$sim(t_a, t_b) = \frac{\sum T_{i,a} T_{j,b}}{\sqrt{\sum (T_{i,a})^2} \sqrt{\sum (T_{i,b})^2}} \quad (3)$$

t_x：トピックx
$T_{x,y}$：トピックxにおけるタグyの特徴量

2.2. 4指標の提案

2.1.1.節~2.1.5.節によって抽出されたトピックに対して流行性と新規性の抽出を行い，シーンに対して要約性と進展性の評価を以降の節で行う．

2.2.1. トピックの評価指標

クラスタごとのシーン数を分析し，トピックiの流行性(trend)の評価を行う．

$$\mathrm{trend}(i) = \frac{\text{クラスタ}i\text{のシーン数}}{\sum_{k}^{all} \frac{\text{クラスタ}k\text{のシーン数}}{\text{全クラスタ数}}} \quad (4)$$

同じ話題をもつトピック群の発生経過日を分析し，トピックiの新規性(origin)の評価を行う．

$$\mathrm{origin}(i) = \frac{1}{\text{新規トピック}i\text{が発生経過日} + 1} \quad (5)$$

†電気通信大学 大学院情報システム学研究科 Gratuate School of Information Systems, University of Electro-Communications

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.

2.2.2. シーンの評価指標

シーンjのトピック内平均tfidf値(以下，TFIDF値と表記)が上位のタグについて，各シーンごとに特徴量を分析し，シーンjの要約性(abst)の評価を行う．重要な情報を多く含み且つ短時間で視聴できるシーンを発見するため，加算した特徴量にa=0.02としたe^{-at}を掛ける．

$$abst(j)=\sum_{j\in TAG(T_x,S_y)\cap k}\frac{tf\cdot idf(TAG_j)}{\sum_{k\in TAG(T_x)}TFIDFf(TAG_k)}\cdot e^{-at} \quad (6)$$

j:トピックx内シーンyにおけるタグ集合
k:トピックxにおけるTFIDF値上位10個のタグ集合
t:シーンjの放送時間量[分]

シーンjが持つタグのうち，当日に発生した新規タグ集合k'に注目し，k'の特徴量を分析する事でシーンjの進展性(progre)を評価する．

$$progre(j)=\frac{\sum_{j\in TAG(T_x,S_y)\cap k'}tfidf(TAG_j)}{\sum_{k'\in TAG(T_x)}TFIDF(TAG_{k'})} \quad (7)$$

j:トピックx内のシーンyにおけるタグ集合
k':トピックxにおけるTFIDF上位20個以内で当日に新規発生したタグ集合

3. 評価実験

3.1. データセット

2009/03/23~03/29と2009/12/17-12/23の二つの期間内にTVのキー局で放送されたニュースシーンに付随するテキスト情報を扱った．表1にデータセットの一部を示す．

表1 :ニュースシーンごとに付随するテキストデータ

ID	放送日	番組ジャンル	ヘッドライン	メモ
10019906	2009/3/29	ニュース/報道	千葉県銚子市長リコール住民投票・失職決定	市立総合病院の診療休止をめぐり，岡野俊昭市長のリコール投票が行われた

3.2. トピック抽出

節 3.1.1.から 3.1.5.までの分析を行い，データセットされた全ニュースシーンについてトピックの抽出を行った．

3.3. 4指標による特性評価

節3.2.で抽出した全トピックと全シーンに対して，各評価指標による分析を行った．

3.3.1. トピックに対する特性評価

12//17~12/23 に放送された主なトピックの trend 値と origin 値を計算し、それらの推移を図2と図3に示す．

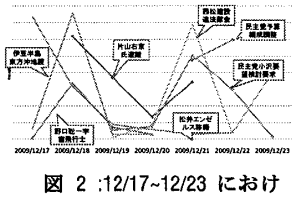

図 2 :12/17~12/23 における trend 値の推移

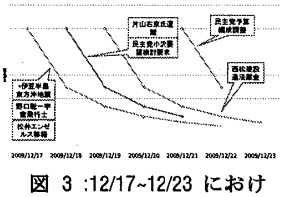

図 3 :12/17~12/23 における origin 値の推移

3.3.2. シーンに対する特性評価

データセットをもとに，各シーンのabst値とprogre値を計算し要約性と進展性の評価を行った．二つの指標によって評価された12/17の「松井エンゼルス移籍」に関するプロ野球トピック内のシーン内容について一部抜粋したものを表4に示す．

表2: 12/17における「松井エンゼルス移籍」に関するトピックのabst値とprogre値の評価

topicID	sceneID	年月日	分類	ヘッドライン	MEMO	abst(j)	progre(j)
20091217_8	1E+07	2009/12/17	スポーツ	【MLB】エンゼルス・松井秀喜・入団会見	松井秀喜がエンゼルスの入団会見を行った。松井秀喜の地元である石川県の様子、エンゼルススタジアムを紹介。エンゼルスと同じアメリカンリーグ西地区に所属するマリナーズにはイチローがおり、松井秀喜との対戦が期待されている。東京墨田区にある大手旅行会社では松井秀喜とイチローの応援ツアー計画の話し合いが行われていた。【会見】松井秀喜【コメント】長谷川滋利氏（元エンゼルス）、マイケソーシア氏（エンゼルス・監督）、羽田勇氏（大手旅行会社）【コメント文】イチロー	2.849237	2.14716
20091217_8	1E+07	2009/12/17	スポーツ	【MLB】松井秀喜・エンゼルス移籍へ	松井秀喜がエンゼルスチームドクターの身体検査を受けた。17日に入団会見を行う。	0.734587	0.392844
20091217_8	1E+07	2009/12/17	スポーツ	<ちゃんスポ！>【MLB】松井秀喜・エンゼルス移籍へ	【出演】中村光宏	0.397736	0.268911

4. 考察

本研究の手法によって抽出したトピックは，付随するテキストデータを参照するに適切な粒度を以てクラスタリングがされているものと考えられる．しかし，同じ話題をTV局や番組によって違う切り口や視点で取材している場合，それらの話題は違うトピックとして処理される傾向にあった．現状では，それらのトピックを違うトピックとして処理をしているが，日毎のトピックの連続性を定義する際に当日の複数のトピックが隣接する日付のある一つのトピックと同一の話題である決定するなら，当日の複数のトピック群を一つのトピックとして見なした方が，後の流行性評価で現状より妥当な数値が与えられる可能性がある．

本研究で提案した指標評価のうち，要約性と進展性の検証に関しては，１つのトピックのメタデータだけに焦点をあてているが，視聴者が動画ニュースシーンを選択する際の評価指標として，今回，要約された内容を含むシーンと進展した情報を含むシーンの抽出が出来たと考えている．本稿で限定的だった検証方法であるメタデータの参照だけに終わることなく，他のトピックに関する検証や実際のビデオデータを使った検証について，引き続き行っていきたい．また，ビデオデータについては，2009/12/16~12/23の期間の録画データを使って，指標によって抽出したシーンを被験者に視聴してもらう事で，検証を終える予定である．

5. 結論

本稿では，テキストデータから同じ内容のシーン集合であるトピックを抽出し，トピックとシーンに対して，4つの特性指標の評価と検証を行った．本研究の評価指標は，大量の動画ニュースリストから見たいシーンを選択するための指標として，評価されている事を検証によって明らかにした．本稿では載せていない動画データの視聴を意識した検証方法や一つだけはなく，より多くのトピックを対象として検証を行う事も引き続き行っていきたい．

関連研究

[1]形態素解析システム Mecab http, //mecab.sourceforge.net/

[2]Newman, M.E.J.(2003),Fast algorithm for detecting community structure in network.

[3] 澤井 里枝, 妹尾 宏, 鹿喰 善明 (2008)， ニュースダイジェスト作成システムにおける重要度算出手法の評価，『情報処理学会研究報告』2008-DBS-144(25), 2008-GN-66(25)

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.