I-048

概念グラフを用いたニュース映像要約手法
The Method to Summarize News Video by Conceptual Graph

林 英俊† 李 龍† 上林 弥彦†
Hidetoshi Hayashi Ryong Lee Yahiko Kambayashi

1. はじめに

通信網の急速な発展により, 近い将来ディジタル放送やインターネットを介してユーザーに膨大な量のデータが高速配信されることが予想される. 配信されるデータ量が膨大になるため, そこから個々のユーザーにとって真に価値のあるコンテンツを発見する事は困難になるであろう.

そこで本研究では, 大量のニュース動画を保存した後, 一定の重要なトピックだけに要約することによって, ユーザーのコンテンツ発見を支援することを目標としている. 動画の要約に関連する研究は多数あるが, ほとんどが音声認識や画像解析技術を中心としたものであり[3][4], ビデオ区画同士の比較はキーワード比較による手法のみにとどまっている. そのため, キーワードのみでなくキーワード間の関連まで含む概念グラフを作成し, それをもとに記事比較を行うニュース要約システムを提案する. 本論では概念グラフを用いて実際に CNN ニュースの要約を実行した.

2. 研究概要

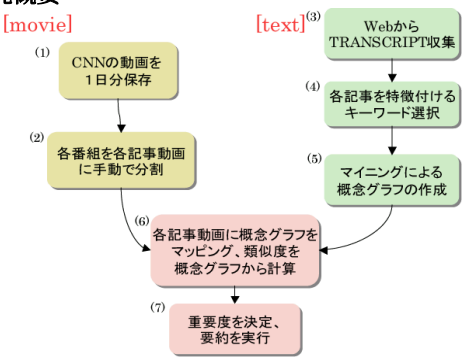

図 1: 要約までのフローチャート

大容量の記憶媒体を最大限に利用して, 従来のアプローチ（Search⇒Store）ではなく, 全ての情報を保存してから検索するアプローチ（Store⇒Search）をとる.

まず, CNN 動画を１日分保存して, 各記事まで分割する{(1), (2)}. 次に動画のメタデータである文字情報を利用して概念グラフを作成する{(3), (4), (5)}. 本研究で提案する概念グラフとは, 単語だけでなく単語間の意味的関連を考えて可視化した図 2 のようなグラフであり, グラフの節点は各キーワードに, 枝は各キーワード間の関連に対応する. CNN は聴覚障害者用に音声情報の写しである CC テキストを TRANSCRIPT として Web 上に公開しており, これをソースとして使用した. さらに, この概念グラフを各記事動画にマッピングする(6). 最後に概念グラフを用いて記事間の類似度計算を行い, 一定の重要トピックに要約する(7).

3. 要約アルゴリズム

3.1. 各記事を特徴付けるキーワード抽出 { 図 1 (4) }

CNN ニュースの音声情報の写しである TRANSCRIPT からストップワードを削除した後, *tf*idf* 法を適用して各記事を特徴付ける重要キーワードを 10 個ずつ抽出する. これは概念グラフの節点の重みを決定する作業である.

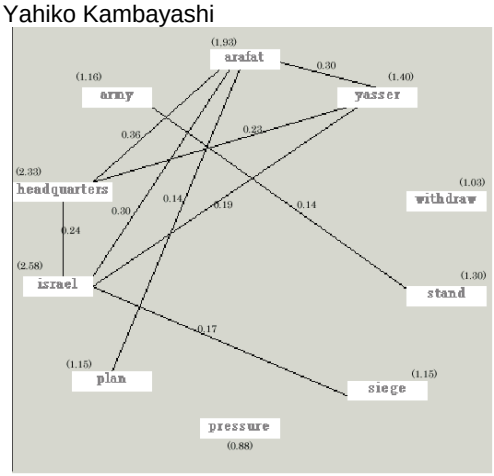

図 2: 概念グラフ

3.2. 各記事の概念グラフ作成 { 図 1 (5) }

キーワード間の関連重みを決定するため TRANSCRIPT から共起関係を抽出する. これにはテキストマイニング技術を用いる[2]. これは概念グラフの枝の重みを決定する作業に相当する.

ある文中にターム t が出現する確率を $P(t)$ とし, 同じ文中に t_{j_1} と t_{j_2} が出現すれば共起と判断すると, Apriori アルゴリズムのパラメータは下式のようになる.

$$Sup(j_1, j_2) = P(t_{j_1} \cap t_{j_2}) \quad Conf(j_1 \Rightarrow j_2) = P(t_{j_2} \mid t_{j_1})$$

この二値から決まる, 記事 D_i 中のターム t_{j_1} と t_{j_2} 間の関連度を示す値 $Ass^i(j_1, j_2)$ を下式のように定義する.

$$Ass^i(j_1, j_2) = Sup(j_1, j_2) \times \frac{Conf(j_1 \Rightarrow j_2) + Conf(j_2 \Rightarrow j_1)}{2}$$

方向性を持たない関連の強さを表したいので, 双方向の Conf 値の平均を取って考えた.

3.3. 各記事を比較して類似記事の発見 { 図 1 (6) }

記事 D_{i_1} と D_{i_2} の類似度（*Similarity*）を計算する. ここでは概念グラフを用いた場合の有効性を評価するために, 従来のキーワード比較による手法と概念グラフによる手法の２通りの方法を用いる. 類似度計算の方法として情報検索の分野で有名なコサイン類似度による手法を用いる.

(1)従来のキーワードによる類似度計算

記事 D_i に出現する単語の *tf*idf* 値を各成分とする特徴ベクトルを $w^i = (w_1^i, w_2^i, ..., w_n^i)$ として, 記事 D_{i_1} と記事 D_{i_2} の間の類似度 $Sim(i_1, i_2)$ はそれぞれの特徴ベクトル w_1^i, w_2^i 間のコサイン類似度として下式のようになる.

$$Sim(i_1, i_2) = \frac{w_1^i \cdot w_2^i}{\sqrt{\left\|w_1^i\right\| \left\|w_2^i\right\|}}$$

(2)概念グラフによる類似度計算

(1)で考えた各単語の *tf*idf* 値を各成分とする特徴ベクトルに加え, 3.2 で求めた *Ass* 値からなる特徴ベクトルも考える. 記事 D_i 中の単語間の *Ass* 値を各成分とする特徴ベクトルを $a^i = (a_{11}^i, a_{12}^i, a_{13}^i, ...a_{n-2,n-1}^i, a_{n-1,n}^i)$ とする. 概念グラフ

† 京都大学情報学研究科社会情報学専攻

による類似度計算では, (1)で定義した w^i とここで新たに定義した a^i の両方のコサイン類似度を加算して類似度を求める. 記事 D_{i_1} と記事 D_{i_2} の間の類似度 $Sim(i_1, i_2)$ はそれぞれの *tf*idf* 値に関する特徴ベクトルを w_1^i, w_2^i, *Ass* 値に関する特徴ベクトルを a_1^i, a_2^i として下式のようになる.

$$Sim(i_1, i_2) = \frac{w_1^i \cdot w_2^i}{\sqrt{|w_1^i||w_2^i|}} + 2 \cdot \frac{a_1^i \cdot a_2^i}{\sqrt{|a_1^i||a_2^i|}}$$

枝の両端の単語が関連するため, Ass 値に関しての類似度に 2 倍の重みをつけた. こうして求めた類似度で記事のクラスタリングを行って類似記事のクラスタを発見する.

3.4. ビデオ要約過程 { 図 1 (7) }

以下の３つの仮定を用いて重要度を決定し, 重要度の高い記事ほど要約動画に含まれるべきであると判断する.

(1)何度も繰り返して放送された内容が重要

繰り返しであるかどうかは類似度で記事をクラスタリングした結果(3.3)から求める. 類似記事の中では最新記事ほど重要であり, 最新記事以外の類似記事は要約に含まずに, 最新記事の重要度をその分高くする. 記事 D_i の類似記事数を $SimArticle(i)$, 総記事数を *NumArticle* として, 繰り返しの仮定に基づく重要度 *imp1(i)*を下式のように定義する.

$$imp1(i) = \frac{\log(SimArticle(i))}{\log(NumArticle)}$$

(2)長時間放映された記事ほど重要

長時間, 放映時間が与えられた記事ほど重要度は高い. 記事 D_i の放映時間を $Duration(i)$, その日の最長記事の放映時間を *MaxDuration* として, 放映時間の仮定に基づく重要度 *imp2(i)*を下式のように定義する.

$$imp2(i) = \frac{\log(Duration(i))}{\log(MaxDuration)}$$

(3)番組のはじめのほうに放送された記事ほど重要

CNN ニュース番組はほとんどが 0 分から始まる 1 時間番組である. よって記事 D_i の放映開始時刻のうち, 分の単位の数字(0～60)を $Minute(i)$ として, 放映開始時刻の仮定に基づく重要度 *imp3(i)*を下式のように定義する.

$$imp3(i) = \frac{\log(60 - Minute(i))}{\log(60)}$$

これら３つの仮定に基づいて記事 D_i の重要度 *imp(i)*を以下のように定義する.

$$imp(i) = imp1(i) + imp2(i) + imp3(i)$$

4. 実験結果･評価

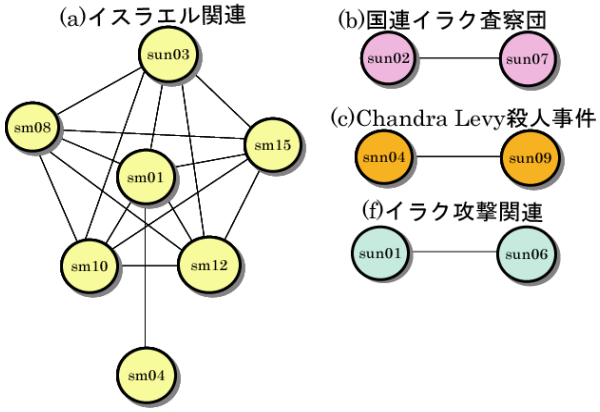

図 3 : 概念グラフによる類似度計算結果

以前の実験でキーワードのみの手法より 概念グラフを用いた類似度計算のほうが有効であった[1]. 図 3 はその類似度計算によって記事をクラスタリングした結果である. この結果を用いて 2002/09/29 のニュースの要約を行った. その結果が図 4 である. この日の 34 記事から imp 値の高いもの 7 つを選び, 2002/09/29 の 30 分要約動画を構成した.

Ranking	Article	imp	imp1	imp2	imp3
1	sun-03	2.84	1.01	0.86	0.96
2	sun-07	2.25	0.39	0.87	0.98
3	sun-06	2.19	0.39	0.81	0.99
4	snn-04	2.10	0.39	0.87	0.83
5	snn-03	1.92	0.00	1.00	0.92
6	sun-05	1.87	0.00	0.95	0.91
7	snn-02	1.83	0.00	0.83	0.99

図 4 : 記事要約結果 (2002/09/29, 全 34 記事)

sun03 は図 3 のクラスタ(a)に含まれ, 7 つの類似記事があるため 3.4(1)の仮定に相当する imp1 値が最大となり, この日の最重要記事となっている. これはイスラエル情勢に関する記事で, この日新しい動きがあったため何度も繰り返し放送されていた. sun07, sun06, snn04 も図 3 のクラスタ(b), (c), (f)に含まれ, それぞれ類似記事が存在するため imp1 が高い値をとり, 重要度が上がっている.

snn03, sun05 はこの日の記事の平均放映時間が約 3 分であったのに対して, 7 分と長時間放映されたため 3.4(2)の仮定に相当する imp2 値が大きくなり要約記事に含まれた.

また snn02 はキューバにハリケーンが今まさに迫っていることを伝える記事である. そのタイムリー性ゆえに 22:00 からの番組のトップ記事として 22:02 と一番早い時間に放映されたため, 3.4(3)の仮定に相当する imp3 値が大きくなり, 要約記事に含まれた.

5. むすび, 今後の課題

4 章で示したようにキーワード間の意味的関連まで考えた概念グラフを用いた比較手法で 1 日分のニュースを要約することに成功した. しかし, ここでの評価は定性的であるため, 引き続き第三者による評価や定量的な評価も行う予定である. その定量的評価を行う際に気づいたことを要約アルゴリズムと[1]で開発中の CNN ニュース要約システムにフィードバックしていくつもりだ.

6. 参考文献

[1] 林英俊, 李龍, 上林弥彦, “概念グラフを用いたニュース映像要約システムの構築”, DEWS2003, 2003.03

[2] Marti A. Hearst, “Untangling Text Data Mining”, ACL ’99, pp.03-10, 1999

[3] 中村裕一, 金出武雄, “ニュース映像からの重要セグメント抽出―画像特徴と言語特徴の相互関係を用いたニュース映像要約―”, 第３回知能情報メディアシンポジウム, pp61-68,1997

[4] Michael A. Smith & Takeo Kanade, “Video Skimming and Characterization through the Combination of Image and Language Understanding”, IEEE, 1997