概念グラフを用いたニュース映像要約システムの構築

林 英俊† 李 龍‡ 上林 弥彦‡

†京都大学工学部情報学科 〒606-8501 京都市左京区吉田本町
‡京都大学情報学研究科社会情報学専攻 〒606-8501 京都市左京区吉田本町
E-mail: †hhayashi@db.soc.i.kyoto-u.ac.jp, ‡{ryong, yahiko}@db.soc.i.kyoto-u.ac.jp

あらまし 記録メディアの大容量化及び動画圧縮技術の発展に伴い長時間の動画データが二次記憶に容易に保存できるようになった. それを利用して本研究では, 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, 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んだ概念グラフを作成し, それをもとに記事の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た場合の比較性能の向上を示し, 実際に CNN ニュースを用いて要約システムを実装した.

キーワード 動画要約, 概念グラフ, テキストマイニング

Construction of News Video Summarization System by Conceptual Graph

Hidetoshi HAYASHI† Ryong LEE‡ and Yahiko KAMBAYASHI‡

†Department of Information Science, Kyoto University Yoshida Honmachi, Sakyo-ku, Kyoto-shi , 606-8501 Japan
‡Department of Social Infomatics, Kyoto University Yoshida Honmachi, Sakyo-ku, Kyoto-shi, 606-8501 Japan
E-mail: †hhayashi@db.soc.i.kyoto-u.ac.jp, ‡{ryong,yahiko}@db.soc.i.Kyoto-u.ac.jp

Abstract The high capacity of storage media and the development of compression technologies for video today have made it easy to store video data for many hours into second storage devices. The goal of this research is to summarize large quantity of news video to a kind of important topics after saving them. While there exist many researches relevant to video summarization, most of them focus mainly on technology of speech recognition or image analysis and compare video segments only by keywords. Therefore, the proposed news summarization system compares news articles based on “Conceptual Graph” formed not only by keywords but also by the connection between keywords. In this paper, we introduce an implemented system using CNN news as an example and show that Conceptual Graph does better in semantic summarization than keywords-only method.

Keyword Video Summarization, Conceptual Graph, Text Mining

1. はじめに

1.1. 研究の動機

通信網の急速な発展により, CD-ROM 1 枚分(650MB のデータ送信が, 光ファイバ(10Mbps)ならば約 9 分, 地上波/BS ディジタル TV 網(21Mbps)ならば約 4 分で可能となった. 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あまりに膨大となるため, 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. 例えば, 多種多様な Web ページから構成されたインターネットの世界では, Google のような検索エンジンが不可欠である. 同様の状況が動画配信の世界でも発生し, 検索エンジンに相当する仕組みとして動画検索や動画要約に対する必要性が高まる事が予想される.

動画要約を行う際に, 類似シーンを発見するために類似度計算を行う過程がある. 動画要約に関するほとんどの研究は音声認識や画像解析技術を中心としたものであり, キーワードのみを動画にマッピングして類似度計算をした後, 要約を実行している.

そこで本研究では, キーワードのみでなくキーワード間の意味的関連まで含む概念グラフを作成して, それを用いて記事間の類似度計算を行ってから, 一定の重要なトピックだけに要約するシステムを構築する. 具体的には図 1 のように, ユーザーが 1 日のスポーツ

ニュースの概要を 30 分で知りたいというクエリを与え，システムは保存されたニュース動画の中から重要なトピックのみを選び出して，30 分の要約動画に編集してからユーザーに与えるシステムである．

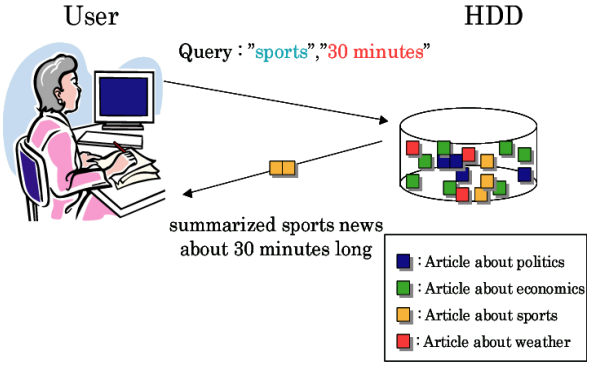

図 1: ユーザーからのクエリ例

1.2. 研究の背景

(1)長時間の動画を保存可能に

近年，記憶媒体の大容量化と動画圧縮技術の進歩により長時間の動画を保存可能となった．現時点でパソコン用でも 160GB といった大容量ハードディスクが登場している．動画圧縮技術も MPEG1，MPEG2，MPEG4 等，多様な高圧縮技術が出現してきており，MPEG1 形式で 1 時間の動画を約 720MB まで圧縮することが可能である．これを上記の 160GB のハードディスクに保存することを考えると，1 つのディスクに 220 時間の動画を保存することが可能となっている．このように大量の動画が保存可能なことを利用して，既に蓄積型テレビという概念が生まれている．これは番組を蓄積して必要な番組だけ都合のいい時間に見るという考え方である．松下，東芝など 4 社は共同で，HDD 搭載の専用受信機を用いて蓄積型双方向テレビサービス"ep"を既に開始している[*1]．SONY もネットワーク接続可能でキーワードを登録すれば電子番組表（EPG）から関連番組を検索して録画する機能を持った<CoCoon>を販売しており，320GB のハードディスクに EP モードで最長 200 時間の録画可能である[*2]．

(2)動画にメタデータとして文字情報が

今後，動画にメタデータとして文字情報が与えられることも多くなるであろう．総務省の発表によると，2003 年に東京・大阪・名古屋で，2006 年には全国すべてのローカル局で地上波放送がディジタル化される予定である．ディジタル化のメリットは多数あるが，その中でもデータ放送が最も社会に影響を与えるであろう．データ放送には XML をベースにしたマルチメディア・コンテンツに対するメタデータ表記方法の国際標準規格である MPEG-7 が用いられる．動画に対するメタデータとしては（タイトル，製作者，製作日時）等が含まれる予定である．また，アメリカでは既に 1979 年に NCI が聴覚障害者用向けの公式文字放送(CC:Closed Caption)を開発しており，現在ほとんどの番組が CC 対応で，CC に対応する CC テキストが公開される場合が多い．CC テキストは，基本的には映像に含まれる音声情報の写しであるが，単なる単語と文の集合ではなく組織化されたテキスト構造を持っているため，うまく解析することにより映像を特徴付ける文字情報として利用可能である．日本でも 2007 年度開始を目指して同様の仕組みを導入することが検討されている．

1.3 研究概要

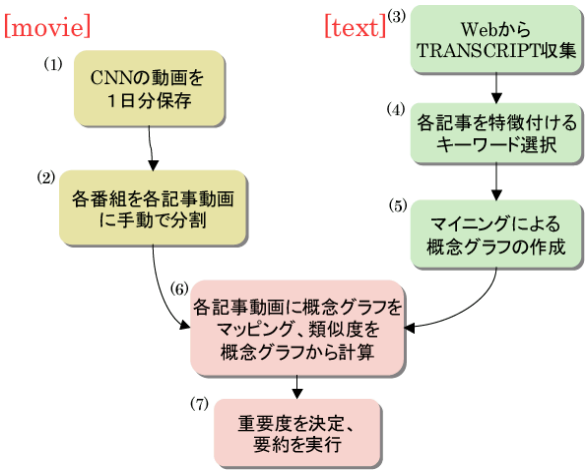

図 2: システム構築のフローチャート

1.2(1)でも述べたように長時間動画が保存可能なことから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．CNN ニュース動画を 1 日分全て保存してから各記事にまで分割する．次に 1.2(2)で述べたように動画にメタデータとして文字情報が入ることを利用して，音声情報の代用となる CC テキストからキーワード間の意味的関連まで考えた概念グラフを作成して，各記事動画にマッピングする．最後に記事間の類似度計算を行って一定の重要トピックだけに要約する．

ドラマや映画は全て見なければおもしろくないが，ニュースやスポーツは一番重要な部分だけ見れば全体がつかめることが多いため要約の効果が高い．さらに CNN ニュースでは CC テキストを TRANSCRIPT として Web 上に公開している．そのため映像ソースとして CNN ニュースを選択した．

実際のシステム構築概要は図 2 のようになる．以後，

[*1] 蓄積型双方向テレビサービス「ep」
http://www.epep.jp/index.html

[*2] ＜CoCoon＞チャンネルサーバ
http://www.sony.jp/products/Consumer/cocoon/

２章でシステム構築のために必要となる基本技術やその関連研究について述べる．３章で図２の各項目の詳細なアルゴリズムについて説明する．４章で CNN ニュースを利用して構築したシステムを紹介する．５章では CNN 記事を利用した実験結果について説明する．

2. 基本的事項，関連研究

2.1. 動画の構造化

映像データのようなマルチメディアデータは，単純にマッチングアルゴリズムが適用可能な文字データとは異なり構造化されていない．従って要約を行うためには映像を意味的に一貫性のある映像区間に分割してそれぞれを特徴付ける必要がある．ニュースやスポーツ映像は比較的わかりやすい構造を持っているため多くの研究がなされており，動画の構造化については他の研究でも様々なアプローチがとられている．

有木[12]は映像をカット点抽出で分割して，音声認識＋テロップ認識によって得られる文字情報から抽出したキーワードで記事映像を構造化する．キーワードから算出した関連度を用いて記事分類を行っている．

Smith, Kanade[13]は音声に関しては TRANSCRIPT の文字情報から音声スキムに分割，キーワード抽出でそれぞれの音声スキムの重要性を決定する．次に映像に関してはカラーヒストグラムを用いたシーン検出によって映像スキムに分割して，カメラモーション，顔やキャプションの認識によりそれぞれの映像スキムの重要性を決定する．最後に音声スキムと映像スキムをさらに高レベルのメタデータを用いてマッピングし，ソース映像の 1/20 以下のビデオスキムを要約映像として作成している．彼らは Infomedia プロジェクト（動画を大量保存してデジタルビデオライブラリを作成，検索・要約を実現）にも携わっている．

中村，金出[15]も Smith ら[13]と同様に映像からはキーフレームを，TRANSCRIPT またはクローズドキャプションから構文解析を用いてキーセンテンスをそれぞれ取り出す．キーフレームとキーセンテンスを動的計画法でマッピングすることで重要セグメントを抽出してから，要約や検索に利用している．

このように動画を構造化するには映像データ＋文字情報（音声データ，クローズドキャプションまたは TRANSCRIPT から抽出）が必要である．本研究では映像の分割とマッピングは手動で行う．しかし従来のアプローチのようにキーワードを動画にマッピングするわけではない．TRANSCRIPT の文字情報からキーワードを選択して，そのキーワードに対してテキストマイニング技術を用いて作成した概念グラフを，各記事動画にマッピングする．

2.2. CNN ニュース TRANSCRIPT

今回は映像ソースとして CNN ニュースの Domestic チャンネルを選んだ．前述のようにアメリカでは既に CC を用いた公共文字放送画始まっている．

現在はデータ放送が行われていないため，メタデータを得ることはできないが，CNN ニュースではニュースの音声情報を TRANSCRIPT として Web 上で公開している[*3]．これをメタデータの代用として利用する．

CNN ニュース TRANSCRIPT のソースコードは単なる音声情報の写しではなく，図３のような簡単な XML のような構造をしており，記事本文以外にも(A)～(E)のような様々な有益な情報を含む．

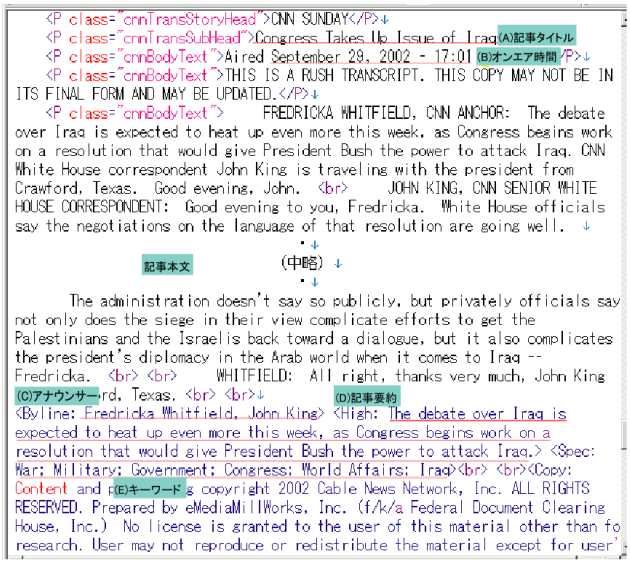

図 3: Web 上の TRANCSCIPT ページのソース

2.3. テキストマイニング

動画のメタデータとして使用する TRANSCRIPT は自然言語であるためデータベース内のデータの様に構造化されておらず，知識を発見して概念グラフを作成するためにテキストマイニングの技術が必要となる．

Hearst[4]や Dixon[5]はテキストマイニングの概念を以下のように定義している．テキストマイニングとは構造化されていないデータから新たに知識を発見する技術であり，従来の技術とは異なったものである．データマイニングは，テキストデータではなく構造化されたデータベースから情報を抽出する．情報検索は新たな知識ではなく文章作成者にとっては既知の情報を提供するのみである．自然言語処理は文法上のパターンを発見するが，知識を得るわけではない．テキストマイニングは，まず自然言語を対象とするため自然言語処理が不可欠であり，さらに情報検索，データマイニングや機械学習の技術を組み合わせて実現される．

[*3] CNN News TRANSCRIPTS
http://www.cnn.com/TRANSCRIPTS/

Feldman[7], Rajman[6]は各単語の品詞を調べ「名詞＋名詞」,「形容詞＋名詞」といった品詞の組み合わせによってターム候補となるものを１つとして扱う. さらにテキストデータからの不必要な品詞（前置詞, 冠詞, 接続詞）の除去, 接辞処理など自然言語処理技術を用いて前処理を行った後, χ^2 分布や相互情報量等の統計的手法を用いて重要タームを決定する. 重要ターム間に相関ルール抽出を行い, その結果を用いてテキスト分類を実行している.

Ahonen[8]も, テキストデータを単語系列データとみなす. 重要な品詞を選択して, 語幹のみに注目した後, 相関ルールの応用である episode rule といった順序パターンを発見している. この episode には順序情報を考慮する serial episode と考慮しない parallel episode があり, 後者を使用すると共起単語セットが得られる. episode rule は mannila[9]が時系列パターンを変形して開発したアルゴリズムで, 下式のようになる.

$knowledge, discorvery, in[4] \Rightarrow databases[5](85\%)$

これは（knowledge, discovery, in）が 4 単語以内に出現した場合,（knowledge, discovery, in, databases）が 5 単語以内に出現する確率が 85%であることを示す.

2.4. 概念グラフ

情報検索やマイニングの結果を可視化するという研究は多くある. WebBrain[*4] や Cat-a-Cone Interface[*5] といった研究は, Business, Companies といったカテゴリと, カテゴリ-サブカテゴリ間の関係からグラフを作成して３次元に可視化することで Web ナビゲーションに利用している.

渡部[10]は ACCENT というテキストマイニングによって導かれた連想関係を可視化するツールを開発している. 相関度を反映させたスプリングで単語間をつないだ仮想的な物理モデルの安定状態を求めることにより各単語のレイアウトを決める技術（スプリング埋め込み技術）を用いて結果を可視化している.

本論で提案する概念グラフ（Conceptual Graph）とは, 単語だけでなく, 単語間の意味的関連を考えた図４のようなグラフであり, 単純に可視化して使用するだけではない. 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し, キーワード間の関連は 2.3 に示したテキストマイニングから求める. 比較要素としてグラフの節点と枝両方の重みを考慮して, 複数グラフ間の類似度計算に用いる. このように意味的関連まで含んだ概念グラフを用いて各記事を比較することで比較性能の向上が望めると考えた.

[*4] WebBrain http://www.webbrain.com
[*5] Cat-a-Cone Interface http://www.sims.berkeley.edu/~hearst/cac-overview.html

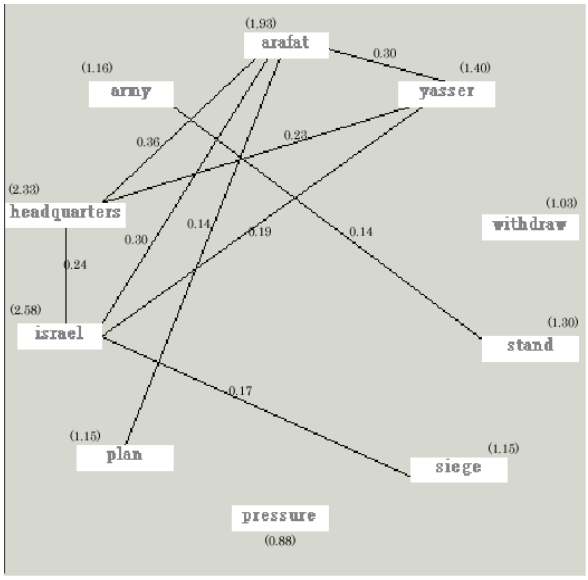

図 4: 概念グラフの例 (sm01)

3. CNN ニュース要約システム

本章では 1.3 の図２中で述べた(1)～(7)各項目の詳細なアルゴリズムについて述べる.

3.1. CNN ニュース動画をハードディスクに保存

2002 年 9 月 29 日の CNN ニュースのうち, ３分程度の記事が続く純粋なニュース番組（CNN Sunday Morning, CNN Sunday, CNN Sunday Night の３番組）をハードディスクに mpeg1 形式でキャプチャした. キャプチャボードから手動で取り込んだが, EPG（電子番組表）と HDD 内臓テレビを使用すれば自動化することは容易であろうと考えられる. これらの番組以外は特集番組や Debate 番組であったため省略した.

3.2. 番組を各記事に分割

この日のニュースは３番組, 合計 34 個の記事から構成されていた. この分割についても手動で行ったが, ニュース映像のように比較的簡単な構造をした動画の分割に関しては, 2.1 で述べたように多くの研究がなされており, これらの画像解析や音声認識の技術を用いれば自動化できる可能性は高いと思われる.

3.3. Web からの TRANSCRIPT ページの収集

CNN の TRANSCRIPT は Web 上のある法則に従ったアドレスに保存されており, 例えば 2002/09/29 の TRANSCRIPT のトップページは下記のようになる.

http://www.cnn.com/TRANSCRIPTS/2002.09.29.html

このページから全記事の TRANSCRIPT へのリンクが出ており, それを辿ることでその日の全ての TRANSCRIPT をダウンロードする.

2.2 の図３のように TRANSCRIPT は簡単な XML のような構造をしている. これを解析することで記事本文のみでなく,（(A)記事タイトル, (B)オンエア時間,

(C)アナウンサー名, (D)記事要約, (E)キーワード, (F)ゲスト名）といった情報を抽出できる.

3.4. 各記事を特徴付けるキーワード抽出

記事中の全単語を対象として概念グラフを作成すると，n 個の単語に対し ${}_nC_2$ 個の関連を抽出する必要があり，膨大なマシンパワーを消費する．よって概念グラフの節点（各キーワード）とその重みを決定するために前処理をして重要単語に絞る.

まず，文章を特徴付ける上で役に立たないストップワード（I, yes, will など）を削除する．ストップワードリストとしては一般的な英語文章を検索対象とする代表的検索システム, SMART システム[*7] で標準的に使われているリストを使用した.

その後, *tf*idf* 法を適用することによっ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．記事 D_i におけるターム t_j の出現頻度を $freq(i,j)$，D_i 中の総ターム数を $TermKind(i)$，ターム t_j が出現する文書数を df_j，記事総数 N として

$$tf_{ij} = \frac{\log(freq(i,j)+1)}{\log(TermKind(i))} \qquad idf_j = \log(\frac{N}{df_j})$$

となる．*tf* 値はタームの網羅性を，*idf* 値はタームの特定性を表現している．メタデータとして抽出したデータのうち，記事タイトル(2.3(A))や記事要約(2.3(D))に含まれる単語はより一層重要であるため，記事本文中の単語より頻度を２倍として *tf* 値を計算した.

記事 D_i におけるターム t_j の重みを $w_j^i = tf_{ij} \times idf_j$ として，この値が大きいものほどその記事において重要な単語であるとみなす．記事 sm01 から *tf*idf* 法によってキーワードを抽出した結果が表 1 である.

表 1：抽出キーワード TOP10 (sm01)

Date	Pro	num	tfidf	word
2002/09/29	sm	01	0.879361	pressure
2002/09/29	sm	01	1.40698	yasser
2002/09/29	sm	01	1.30062	stand
2002/09/29	sm	01	1.165	army
2002/09/29	sm	01	1.15262	siege
2002/09/29	sm	01	1.15262	Plan
2002/09/29	sm	01	1.03313	withdraw
2002/09/29	sm	01	2.58489	israel
2002/09/29	sm	01	2.33001	headquarters
2002/09/29	sm	01	1.93459	arafat

3.5. 各記事の概念グラフ作成

概念グラフの枝（各キーワード間関連）の重みを決定するために TRANSCRIPT という自然言語テキストから共起関係を抽出する. 2.3 でも述べたように様々なテキストマイニング手法が考えられる.

この要約システムでは 3.4 で抽出されたキーワードに対して，相関ルール抽出の代表的アルゴリズムである Apriori アルゴリズムを応用して適用する．そして *israel* ⇔ *siege* のように，相関の左辺と右辺をひとつの単語に限定した双方向の相関ルールを求める.

ある文中にターム t が出現する確率を $P(t)$ とし，同じ文中にターム t_{j_1} と t_{j_2} が出現すれば共起と判断すると, Apriori の２つのパラメータは下式のようになる.

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

単語間の関連度を示す値として，この２つのパラメータから総合的に決まる値 *Ass* 値を定義する．Apriori アルゴリズムの *Conf* は上式が示すように方向性を持つ．今回は２つの単語間の関連を考えたいので，双方向の *Conf* の平均をとって考えた．記事 D_i 中のターム t_{j_1} と t_{j_2} 間の $Ass^i(j_1,j_2)$ を下式のように定義する.

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

3.4 で記事 sm01 から抽出された 10 個のキーワード間に全てに *Ass* 値を考える. 2.4 にあった図 4 は概念グラフを作成した後，*Ass* 値が一定の閾値を越えた重要な枝に絞って可視化したものである.

3.6. 各記事を比較して類似記事の発見

記事 D_{i_1} と記事 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.

２つの文章間の類似度を計算するとき，情報検索の分野ではコサイン類似度による類似度計算を行うことが多い．これは，まず各文章を特徴ベクトルで表現して，２つの特徴ベクトルがなす角度のコサインで類似度を計算する手法である．この類似度は２つの特徴ベクトル間の角度を θ とした時の $\cos\theta$ の値となっており，区間[0,1]の値をとるように正規化されている.

(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 $w^i = (w_1^i, w_2^i, \ldots\ldots, w_n^i)$ 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表現される.

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

(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，3.5 で求めた単語間の関連を示す *Ass*

[*7] SMART（Salton & McGill, 1983; Salton, 1988）

値から構成される特徴ベクトルについても考える．記事 D_i に出現する単語間の *Ass* 値を各成分とする特徴ベクトルを $a^i = (a^i_{11}, a^i_{12}, a^i_{13}, \ldots\ldots a^i_{n-2,n-1}, a^i_{n-1,n})$ とする．概念グラフ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 の両方について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^i_1, w^i_2，*Ass* 値に関する特徴ベクトルを a^i_1, a^i_2 として両方のコサイン類似度を考えて下式のようになる．

Ass 値からなる特徴ベクトルのコサイン類似度には２倍の重みをつけて計算した．

$$Sim(i_1, i_2) = \frac{w^i_1 \cdot w^i_2}{\sqrt{|w^i_1||w^i_2|}} + 2 \cdot \frac{a^i_1 \cdot a^i_2}{\sqrt{|a^i_1||a^i_2|}}$$

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．2002/09/29 の CNN News に対して(1), (2)の２通りの方法で実際に実験を行った．実験結果は 5.1 に示してある．

3.7. ビデオ要約の過程

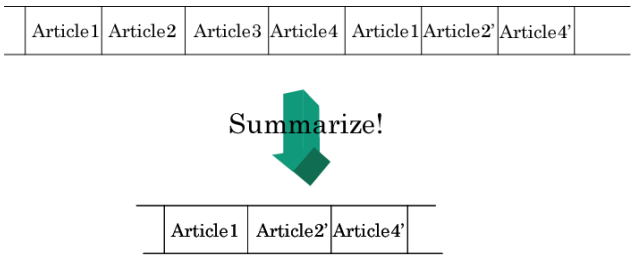

図 5: 要約の基本的な考え方

(1)要約アルゴリズム

その日における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．重要度を決定する基準として以下の３つの仮定を用いる．

- **何度も繰り返して放送された内容が重要**

  内容が同一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(3.6)から求める．類似記事の中では後で放映された記事ほど新しい情報を含む最新記事と判断できるため重要である．

  記事 D_i の類似記事数を $SimArticle(i)$，その日の総記事数を *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．

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

- **長時間放映された記事ほど重要**

  ニュース作成者は各記事に貴重な放映時間を割くわけであるから，当然長時間の枠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．

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

- **番組のはじめのほうに放送された記事ほど重要**

  例えば 17:00 からの典型的な１時間ニュースを考えてみる．17:00 に放映された記事は間違いなくトップ記事であろうし，その後 17:15 くらいまで重要記事が並び，17:30 くらいに専門家を交えて討論などが始まったりする．17:50 くらいには重要度が低い天気予報や次回予告が放送されることが多い．

  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．繰り返しの仮定が最も重要度に影響するように 2:1:1 の重みをつけて加算する．*imp1, imp2, imp3* は区間[0,1]の値をとるように正規化されているため，*imp(i)*は区間[0,4]の値をとる．

$$imp(i) = 2 * imp1(i) + imp2(i) + imp3(i)$$

(2)カテゴリキーワード

要約の際にユーザーはカテゴリキーワードと要約動画の長さをクエリとしてシステムに与える．各記事の TRANSCRIPT には CNN によってキーワード(2.2.(E))が与えられており，記事分類が行われている．ここに出現するキーワード集合からリストを作成して，その中からユーザーにカテゴリキーワードを選択させる統制キーワード方式を採用する．

4. システム実装

本章では３章で述べたアルゴリズムに基づいて実際に開発した CNN ニュース要約システムを紹介する．

(1)TRANSCRIPT 収集，(2)キーワード抽出，(3)概念グラフ作成，(4)類似度計算の各種ボタンを押すとそれぞれの処理を実行して，要約の前処理が完了する．

最後にユーザーがクエリとしてカテゴリキーワードと要約動画の長さをクエリとして与えて(5)要約実行ボタンを押すと，Movie 部分に要約動画を再生して，記事部分には要約動画に含まれる記事が並ぶ．

開発言語としては，インターフェイス部分に Visual Basic をプログラム部分には Perl と MySQL を用いた．

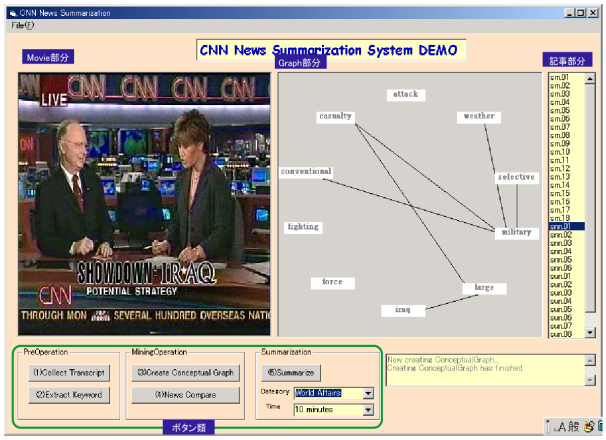

図 6: CNN 要約システムインターフェイス

5. 実験結果，評価

5.1. 比較性能評価

概念グラフを用いた類似度計算の有効性を示すために２通りの手法で計算して，結果を評価する．

(1)キーワードのみによる類似度計算（従来の手法）

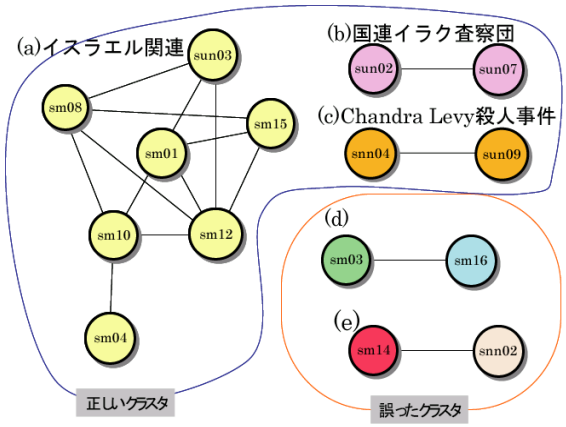

図 7: キーワードのみの類似度計算結果

図７はキーワードのみを用いて計算した類似度で記事をクラスタリングした結果である．各クラスタについて，

(a) Israelis pull back from Yasser Arafat's Compound

これはイスラエル軍がブッシュ大統領の声明を受けて，アラファト議長の屋敷の包囲を解いて撤退をはじめた記事である．この日最も重要であったため朝から晩までで７回も放送されている．

(b) U.N., Iraqi Officials Meet in Austria

sun02, sun07 共に国連のイラク査察団がウィーンに集まり，イラクの高官と査察の方法について議論することを伝える記事．sun02 が第一報でありロンドンからの中継で以前のイラクの対応が中心の記事であるのに対して，sun07 はウィーン（現地）からの続報であり会議が近づいているため実況中継が中心である．

(c) New development in Chandra Levy investigation

sun09, snn04 共に Chandra Levy 殺人事件についての続報．sun09 は FBI のプロファイラを交えて軽く事実を伝えており，snn04 は CNN の捜査アナリストを交えてより詳細に事件について分析している．

これら３つのクラスタはうまく抽出できたのであるが，(d)と(e)は比較を誤ったクラスタである．

(d) 共通 Keyword=”teach”

(sm03) 今日の子供たちに作法を教えるべきだ
(sm16) ジョージア州が霊魂創造説を公立学校で教えることを認めた
いずれも”教える” 事に関する記事であるため”teach”というキーワードがかなり上位キーワードとなっており，そのため *Sim* 値が高くなってしまったが，実際には何の関連もない記事である．

(e) 共通 Keyword=”Cuba, Havana”

(sm14) アメリカ食品をキューバ人が好んでいる
(snn02) ハリケーン Lily がキューバに迫っている
共にキューバに関する記事で，Cuba, Havana という共通キーワードのうち，キーワード”Cuba”がどちらの記事にとってもかなり重要キーワードとなっており，そのため *Sim* 値が高くなってしまった．この２つの記事にもほとんど関連は無い．

(2)概念グラフによる類似度計算

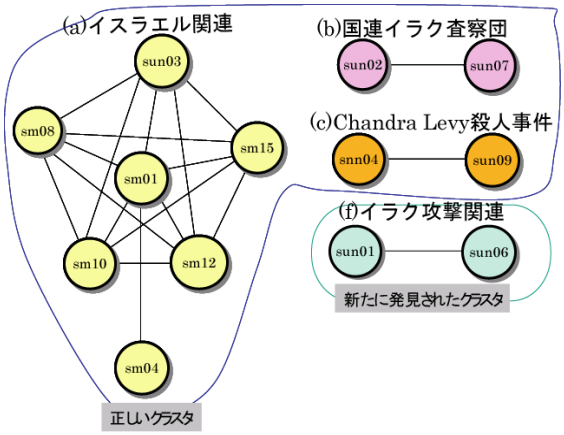

図 8: 概念グラフによる類似度計算結果

図８は概念グラフを用いて計算した類似度で記事をクラスタリングした結果である．(a), (b), (c)はキーワードのみの時と同じクラスタである．5.1 と比べ，(d), (e)といった間違ったクラスタが除かれ，それに加えて(f)という新たな類似記事が発見されており，比較性能が向上している．

(f) Congress Takes Up Issue of Iraq

sun01, sun06 共にアメリカ国会がイラク攻撃に関して議論していることを伝えている．どちらも John King による報道＋民主党反対派の発言＋共和党賛成

派の反論＋英国首相ブレアの発言という構成をしているが, sun01 だけ後半 1/3 にイスラエル情勢を織り交ぜて記事を展開しているため共通部分は前半 2/3 となっている. 概念グラフによる比較で半分以上一致した記事を類似記事と判定することができた.

5.2. 要約結果

表 2：要約結果（2002/09/29; カテゴリ選択なし, 30 分）

Article	Title	imp
sun03	Israelis lift siege of Afarat's compound	2.84
sun07	U.N., Iraqi officials meet in Austria	2.26
sun06	U.S. lawmakers consider Iraq question	2.20
snn04	New development in Chandra Levy investigation	2.10
sun03	Interview with Rob Sobhani	1.92
sun05	Interview with Howard Shua-Eoan	1.87
snn02	Hurricane Lily threatens Cuba	1.83

5.1(2)で概念グラフを用いて計算した比較結果を利用して,繰り返しの仮定に基づく重要度 *imp1(i)*を計算する. 3.7 にも述べたように, 計算した *imp1(I)*と, *imp2(i),imp3(i)*から *imp(i)*を計算して各記事の重要度を決定する.

図８のクラスタリング結果に基づいて各記事の重要度を決定して, 2002/09/29 の CNN ニュースを(カテゴリ選択なし, 30 分)というクエリで要約した結果が表２である. sun03 は 5.1(2)のクラスタ(a)に含まれ, 7 つの類似記事があるため最重要となっている. sun07, sun06, snn04 も 5.1(2)のクラスタ(b), (c), (f)に含まれ, 類似記事が存在するため重要となっている. sun03, sun05 は約 7 分と長時間放映されたため, また snn02 は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それぞれ重要となっている.

6. むすび, 今後の課題

5.1 で示したように記事の比較の際に従来のキーワードのみの手法よりも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手法のほうがよい実験結果が出たため, 概念グラフを用いたニュース記事比較の有効性を示せた. 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. 今後の課題としては以下のものがある.

まず, 情報検索からのアプローチとしてはシソーラスを用いた同義語, 類義語処理が考えられる.

自然言語処理からのアプローチとしては, 接辞処理, 品詞分析, 連語処理がある. 英語には play, plays, played のように用いる構文によって同じ概念をどのような語形で表現するかが変化する. 接辞処理はあらかじめ用意された規則に従って接尾辞を繰り返し削除, 語基を残すことである. 次に, 接辞処理を行った単語を辞書の見出し文字列と一致させることで各語に品詞を決定できる. 最後に, 連語処理は各語に与えられた品詞を利用して特定の構造をした名詞句を抽出する, または統計情報に基づいて句を定義することでキーワードではなく white house, Yessar Arafat といったキータームで文章を特徴付けることができる. ターム単位で概念グラフを作成する予定だ.

データマイニングからのアプローチとしては 2.3 に述べたように相関ルールの変わりに episode rule を用いた概念グラフ作成が考えられる.

ニュースの特性を生かしたアプローチも考えられる. ビデオクリップの開始, 終了位置が TRANSCRIPT に明示してあるため, クリップ直前または直後のアナウンサーの発言（クリップの要約である可能性大）に注目する. また通常キーワードとして名詞が重要であるが, 動画は動作を含むため動詞キーワードのほうがうまく特徴づけできる可能性がある.

文　献

[1] 西尾章次郎, 田中克巳, 上原邦昭, 有木康夫, 加藤俊一, 河野浩之, “岩波講座 マルチメディア情報学<８> 情報の構造化と検索”, 岩波書店, 2000

[2] Jiawei Han & Micheline Kamber, “Data Mining”, morgan kaufmann publishers, 2000

[3] 徳永健伸, “言語と計算５ 情報検索と言語処理”, pp.11-35, 東京大学出版会, 1999

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

[5] Mark Dixon, “An Overview of Document Mining Technology”, October 4, 1997

[6] Rajman M. and Besancon R., “Text Mining: Natural Language techniques and Text Mining applications”, Proceedings of DS-7, Oct7-10, 1997

[7] Feldman R., Fresco M., Yakkov K., Lindell Y., Liphstat O., Rajman M., Schler Y., and Zamir O., “Text Mining at the Term Level”, PKDD ’98, 1998

[8] Ahonen H., Heinonen O., Klemettinen M., and Verlcamo I., “Applying Data Mining Techniques in Text Analysis”, Report C-1997-23, Department Of Computer Science, University of Helsinki, 1997

[9] Heikki Mannila, “Data mining: machine learning, statistics, and databases”, In Proceedings of the 8th International Conference on Sientific and Statistical Database Management, pp1-6, Sweden, 1996

[10] 渡部勇, “ビジュアルテキストマイニング”, 人工知能学会誌16巻2号, pp226-232, 2001.3

[11] 有木康夫, “マルチメディア情報の解析と統合”, 人工知能学会情報統合研究会, 2000.11

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

[13] 中村裕一, 外村佳伸, “見たい部分を短時間で”, 電子情報通信学会誌 Vol.82 No.4, pp.346-353,1999

[14] 中村裕一, 金出康雄, “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