テキストセグメンテーションを用いたマルチメディア検索システム

髙橋 雅仁* 坂田 茂** 西本 由之** 古屋 大亀* 森元 逞***
*久留米工業大学 **㈱ジェイ・フィット ***福岡大学工学部

1．はじめに

文章を段落などの話題の単位で自動的に分割するテキストセグメンテーション技術は、全文検索・要約などの情報検索関連分野、文章構造の解析などへの幅広い応用が期待できる。本研究では、この技術をマルチメディア検索の分野に応用することを目指す。具体的には、テレビニュースや映画、講義ビデオなどのマルチメディアコンテンツを蓄積したデータベースを検索し、ユーザが欲する話題に関連する一連の場面を切り取って出力するシステムの開発を目指す。

2．本研究の背景

（1）ブロードバンドの普及

ADSL、CATV、光ファイバーなどのブロードバンド回線の普及が進んでいる。『平成14年情報通信白書』によれば、2005年度末には、高速・超高速インターネットが1,977万世帯に普及すると予測されている。

（2）動画コンテンツの検索手段の要求

ブロードバンドの普及に伴い、映画やeラーニング教材などの動画を含むマルチメディアコンテンツが増大し、これらのコンテンツの中から、自分が必要とする一場面だけを見たいというユーザの要求が高まることが予想される。

3．本研究の課題

従来の動画を含むマルチメディアコンテンツの検索技術は、以下の問題点を有している。

（1）マルチメディアデータの加工が必要

検索対象とするマルチメディアデータに対して、予め、話題毎にデータを分割し、かつ、それらの分割された各データに対して、キーワードを付与しなければならない。この作業に多大の労力とコストを要する。

（2）検索もれの発生

ユーザが入力したキーワードと、分割されたマルチメディアデータの各々に付与されたキーワードとの照合により検索を行う。そのため、同じ意味内容でもキーワードが一致しない場合は検索もれが発生する。

本研究は、意味ネットワークを用いた文脈情報の生成手法、および、それに基づくテキストセグメンテーション手法を用いて上記の問題点を改善することを目指す。

本方式は次の特長をもつ。

（1）テキストセグメンテーション処理を行うことにより、マルチメディアデータを分割する作業が不要となる。

（2）ユーザが入力した質問文（または、キーワード）と、分割済みの各マルチメディアデータに含まれる語彙と

の意味的な類似度を意味ネットワーク上の単語の活性状態を参照して判定することにより、マルチメディアデータへのキーワード付与作業が不要となる。また、検索もれを軽減することが可能となる。

4．テキストセグメンテーションを用いたマルチメディア検索の方法

マルチメディアデータから取り出したテキストデータをテキストセグメンテーション処理を行って話題の単位で分割することにより、ユーザの質問文と意味的に関連性が高いマルチメディアデータの一部分を上記の分割されたテキスト単位で出力することが可能となる。処理手順は次の通りである（図１）。

（1）検索対象となるマルチメディアデータ（動画データ＋音声データ）に含まれる音声データを音声認識ソフトを用いてテキストデータ化する。

（2）（1）で得たテキストデータに対して、テキストセグメンテーションを行い、話題の単位で分割を行う。また、分割位置を動画データに関連付けて格納する。

（3）ユーザが質問文を入力すると、質問文と分割された各部分テキストとの意味的な類似度を求める。

（4）類似度がもっとも高い部分テキストに対応するマルチメディアデータの再生開始、終了位置を決定し、再生を行う。

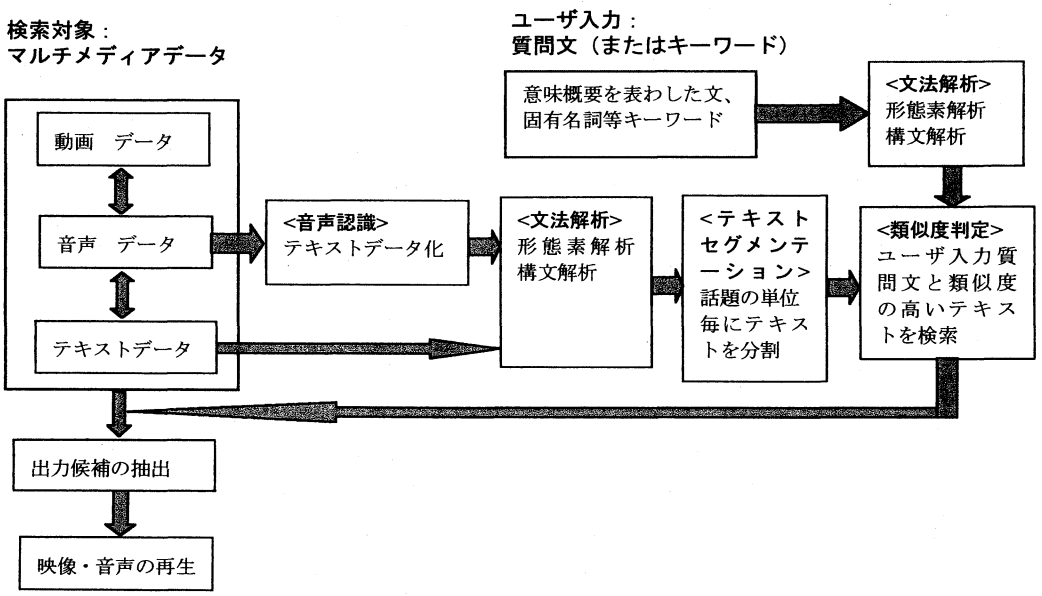


図１　マルチメディア検索システムの処理の流れ

5．話題境界認定に関する基礎実験

（1）実験用テキストデータ

・毎日新聞 1991 年の記事データから内容の異なる記事を７項目つないだデータ

（2）形態素解析

上記の実験用テキストデータに対して、形態素解析ツール「JUMAN」（京都大学）を用いて形態素解析を行った。さらに、形態素解析における単語分割の誤りには手作業で修正を施した（例：「１」「日」「付け」→「１日」「付け」）。手直しをしたファイルとしていないファイルの２種類を実験用テキストデータとした。

（3）テキストセグメンテーション

「Hearst 法」と我々が開発中の「意味ネットワーク上の単語の活性度の変化を用いた方法[1]」の２つの方法で精度を比較し、適合率と再現率を求めた。

適合率＝γ／β×100(%)

再現率＝γ／α×100(%)

αは、話題境界数

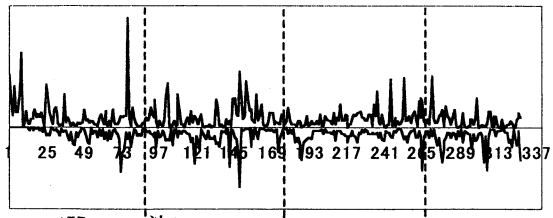
βは、プログラムが出力した話題境界数

γは、プログラムが出力した話題境界中の正解数

図２に２つの方法による実験結果のグラフを示す。図２において、横軸は入力テキストの各単語位置を示し、破線で記した縦軸は話題境界を示している。

図２の上段のグラフにおいて、極大点は話題の開始を表し、極小点は話題の終了を表す。また、下段のグラフにおいて、極小点は話題境界を表す。グラフより、「Hearst 法」では段落境界がほぼ捉えられているが、「単語の活性度の変化を用いた方法」では２番目の段落境界位置が捉えられていないことがわかる。

<単語の活性度の変化を用いた方法>



<Hearst 法>

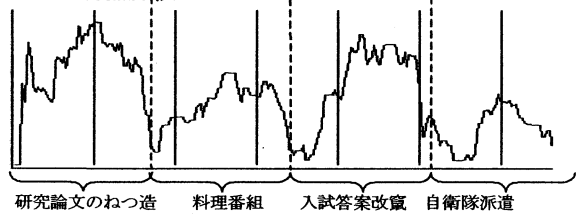


図２　話題境界認定の実験結果

続いて、表１に各テキストデータに対する話題境界認定精度を示す。

表１　話題境界認定精度

	Hearst 法		活性度を用いた方法	
手直し	なし	あり	なし	あり
適合率(%)	100.0	100.0	66.6	50.0
再現率(%)	100.0	100.0	66.6	50.0

6．検索システムプロトタイプの開発

マルチメディア検索システムのプロトタイプの開発を行った。以下に成果の概要を報告する。

（1）実験用データ

テレビニュース 11 本（ニュース 34 項目を含む）の録画データ。

（2）主要ソフト

・音声認識　「Julius」（京都大学）

・テキストセグメンテーション（単語の活性度の変化を用いた方法）

テキストセグメンテーションについては、

基礎実験の結果を踏まえ、意味ネットワークを生成するために必要な名詞と動詞の組（例：絵-を-描く）約20万件を格納した共起辞書を日本経済新聞1年分の記事を用いて作り直し、品質改善を図った。特に、複合名詞を1単位として登録するようにした（例：狂・牛・病・問題-で-揺れる）。

（3）実験結果

・音声認識の精度（名詞、動詞の認識率）
  73.9%（標準設定時）
  83.3%（音声認識パラメータ調整後）
・テキストセグメンテーションの精度
  適合率：55.2%
  再現率：83.6%
・ニュース項目の検索精度
  第1候補の映像の正解率：72.0%

図3に本システムのキーワード入力画面、図4に検索結果の出力画面を示す。図4において検索結果の候補の中から見たいファイルを選択すると、映像が再生される。

7．おわりに

テキストセグメンテーションを用いたマルチメディア検索システムについて報告を行った。音声認識がある程度の品質で行える場合には、ニュース等の話題が明確に分かれたコンテンツに対して実用的な検索システムを開発できる可能性が示された。

<謝辞>

本研究の一部は、2002年度の中小企業総合事業団「課題対応新技術研究調査事業」の委託開発テーマとして実施した。また、本研究で利用させていただいた JUMAN、Julius の開発関係者、新聞記事データを使用させていただいた毎日新聞社、日本経済新聞社、ニュース映像を利用させていただいたｱｰﾙ･ｹｰ･ﾋﾞｰ毎日放送に感謝致します。

<参考文献>

[1] 高橋，森澤，吉村，首藤：「キーワードの活性度の変化を用いたテキストセグメンテーション」，2000 年情報学シンポジウム論文集，pp. 145～152，2000.

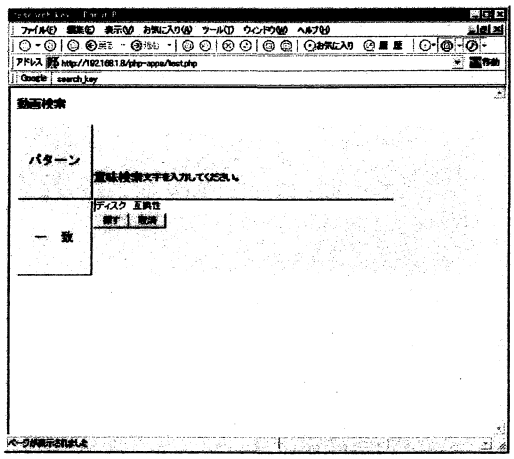


図3　キーワード入力画面

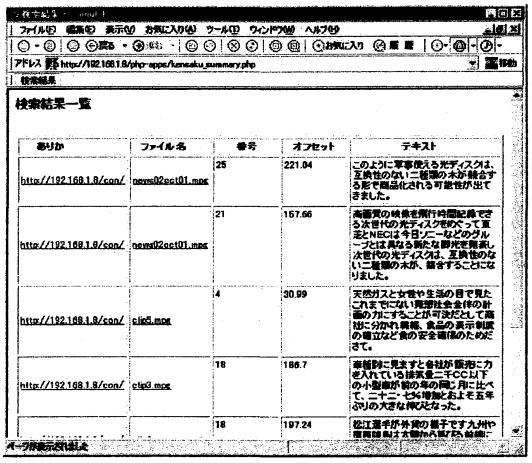


図4　検索結果出力画面