D_028

視聴者の好みとニュースの重要度を考慮したダイジェストの作成

Creation of News Digest Considering both the User's Preference and the News' Importance

澤井 里枝†, 沼田 誠†, 松村 欣司†, 上野 幹大†, 金次 保明†, 八木 伸行†
Rie SAWAI, Makoto NUMATA, Kinji MATSUMURA, Mikihiro UENO,
Yasuaki KANATSUGU, Nobuyuki YAGI

1 はじめに

従来の放送サービスでは，ある特定の視聴者モデルを想定し，あらゆる視聴者に対して同じコンテンツを提供してきた．しかし，受信機，周辺状況，視聴者の嗜好・理解度・障害状況などの視聴環境がますます多様化しているために，個々のニーズに対応した放送サービスへの要望が高まっている．そこで筆者らは，さまざまな視聴環境に適応してコンテンツを自動変換・提示する放送システム AdapTV の研究を行っている [1]．AdapTV では，コンテンツに付加されたメタデータと，視聴環境を記述した視聴環境プロファイルからコンテンツを動的に変換することにより，一つのコンテンツでさまざまな視聴環境に合わせたサービスを提供することを目指す．

本研究では，AdapTV の枠組みの中でニュース番組をコンテンツの対象とし，複数のニュース番組から必要なトピックのみを抽出して必要な順序に並べたダイジェストを作成する．これにより，視聴者は本日のニュースを 10 分で見たり，今週一週間のニュースを 20 分で見るといったダイジェスト視聴が可能となる．従来のダイジェスト作成システムでは，いかにユーザの嗜好を正確に学習するか，どれだけユーザの嗜好に合ったトピックを抽出するか，といったことが主なテーマであった．しかし，ユーザの嗜好に合ったニュースを提示するだけでは，ユーザは現在話題となっている情報や緊急情報などの重要なニュースを見逃してしまう恐れがある．そこで本研究では，視聴者の好みとニュースの重要性の双方に対して適応的にダイジェストを自動作成するための手法について検討し，その試作システムを実装した．

2 ダイジェスト作成システム

本システムの概要を図 1 に示す．本システムでは，ニュース番組の映像と音声，それに付随する字幕を収録する．一般に，放送局はより視聴者に伝えたいニュースや緊急性の高いニュースを番組の冒頭で伝えたり，複数のニュース番組で何度も伝える．これを利用して，本システムはニュースの内容を解析した後，放送状況により重要度を計算する．重要度とは，放送局がそのニュースを伝えたい度合い，つまり放送局の意図を意味することとする．また，本システムでは，ニュース番組が放送されていない時間帯も含めてデータ放送を常時収録しておく．番組だけでなくデータ放送の放送状況も複合的に活用することで，重要度の細かな変化や新規ニュースの出現を検知する．

一方，視聴者側では，ニュース番組とデータ放送の視聴履歴を蓄積する．蓄積する情報には，視聴したトピックと視聴時間，早送り・巻き戻し・スキップ・スロー再生といったリモコン操作を含む．一般的に，視聴者は興味のあるトピックを視聴し，録画番組の視聴中であればあまり興味のないトピックに対してスキップや早送り操作を行う．また，データ放送の閲覧においても，ニュース一覧から興味のある項目のみを選択して中身を読む．そこで，本システムは，視聴者の嗜好を視聴者の番組視聴履歴やデータ放送の閲覧履歴から自動的に学習し，視聴者の嗜好を表現したプロファイルを作成する．次に，前もって解析しておいたニュースの内容とプロファイルを比較して類似度を算出する．

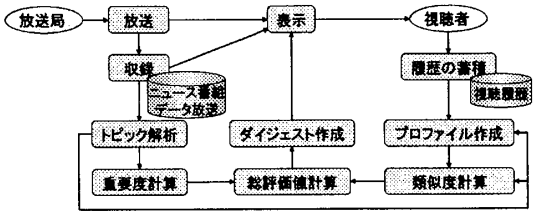

図 1: システムの概要

類似度とは，トピックの内容とプロファイルがマッチしている度合いを意味するものとする．

最後に，求まった重要度と類似度から総評価値を計算してダイジェストを作成する．

以下，2.1 節でトピックの解析，2.2 節で類似度の計算，2.3 節で重要度の計算，2.4 節で総評価値の計算とダイジェストの作成について述べる．

2.1 トピックの解析

ニュース番組は，国会の模様，野球の試合結果，気象予報といったさまざまなトピックから構成されており，本研究ではそれらのトピックの切れ目がメタデータとして番組に付加されていることを想定する．トピックの内容を解析するために，まずニュース番組の字幕の形態素解析を行う．形態素解析とは，言語で意味をもつ最小単位の語に分解して品詞を判別するものであり，本システムでは茶筌 [2] を用いた．また，データ放送コンテンツはダイジェストの構成に直接必要ではないが，本システムでは重要度の計算やプロファイルの作成に用いるため，同様に解析する．

次に，*tf-idf* 手法を用い，全ての語に対して重みを計算する．*tf-idf* 手法とは，該当トピックでの出現頻度が高く，他のトピックでの出現頻度が低いもの，つまりそのトピックの内容をより特徴付ける語に対して高い評価値を与える手法であり，*tf*(Term Frequency) 値と *idf*(Inverse Document Frequency) 値を tf_{ij} = トピック d_i中の語 k_jの出現頻度, $idf_j = log(1/$全トピック中 k_j を含むトピックの割合$) + 1$ と定義する．そして，あるトピック d_i における語 k_j の重みを $w_{ij} = tf_{ij} \times idf_j$ とする．

2.2 類似度の計算

ニュース番組やデータ放送の視聴履歴から視聴者のプロファイルを以下のように作成する．

語の総数を n とし，トピック d_i のコンテンツベクトル D_i を

$$D_i =< (k_1, w_{i1}), ..., (k_n, w_{in}) >$$

と定義する．一方，ユーザ q_u のプロファイルベクトル Q_u を，ユーザがニュース番組やデータ放送を視聴するたびに，次のように更新する．

$$Q_u^{'} = Q_u + aD_i =< (k_1, z_{u1}), ..., (k_n, z_{un}) >$$

ここで，$Q_u, Q_u^{'}$ は更新前，および更新後のプロファイルベクトル，a は視聴したトピックの内容を考慮する度合いを表すパラメータ，z_{uj} はユーザ q_u の語 k_j に対する重みである．パラメータ a は以下のルールに従った値とする．

†NHK 放送技術研究所

- ニュース番組あるいはダイジェスト，データ放送の視聴中にあるトピックを視聴した場合，$a > 0$ とする．
- ニュース番組やダイジェストの視聴中にあるトピックのスキップ再生や早送りをした場合，あるいはデータ放送の閲覧中にニュース一覧から選択しなかった場合，$a < 0$ とする．

ユーザ q_u に対するトピック d_i の類似度は，

$$Similarity_{ui} = \Sigma_k(w_{ik} \times z_{uk})$$

とする．$Similarity_{ui}$ の値が大きいトピックほど，そのトピックに対するユーザの関心度が高いことを意味する．

2.3 重要度の計算

放送状況から各トピックの重要度を計算する．重要度は複数のルール r を組み合わせて算出する．以下にその一部を示す．

- トピックの放送順序について，番組のより冒頭に放送されるトピックの方が重要度は高い．
  番組中に放送されるトピック数を n，あるトピック d_i の放送順位を l_i とすると，トピック d_i の重要度を n/l_i とする．
- 12 時のニュース，19 時のニュース，21 時のニュースなど，より多くの番組で報道されるトピックの重要度は高い．
  トピック d_i, d_j 間の類似度を S_{ij}，トピック d_i, d_j が放送された時刻の差を T_{ij} とすると，$\Sigma_j \frac{S_{ij}}{\sqrt{T_{ij}}}$ とする．
- データ放送でより上位に位置し，かつ長時間，頻繁に放送されているトピックほど重要度は高い．
  項目数を n，あるトピック d_i に関して表示順位 l_{ip} のときの放送時間を t_{ip} とすると，$\Sigma_p\{\sqrt{t_{ip}} \times n/l_{ip}\}$ とする．

それぞれのルールについて各トピックの重要度を決定する．ルール r_j から得られたトピック d_i の重要度を v_{ij} とすると，トピック d_i の総合的な重要度 $Value_i$ を

$$Value_i = \Sigma_j(m_j \times v_{ij})$$

と定義する．ただし，m_j はルール r_j を考慮する度合いを意味し，ユーザの視聴形態から決定する．

2.4 ダイジェストの作成

前節までに算出した類似度と重要度を総合的に評価してダイジェストを作成する．ただし，ユーザが希望する視聴時間に合わせて，ダイジェストの総時間が視聴時間を超えないようにする．

ユーザ q_u に対するトピック d_i の総評価値 E_{ui} を，

$$E_{ui} = x \times Similarity_{ui} + y \times Value_i \quad (1)$$

とする．ただし，x, y は正数であり，類似度と重要度のどちらをより重要視するかに応じて決定する．ダイジェストの再生順序は総評価値が高い順とする．また，トピック間の類似度が一定以上のものを関連するトピックとみなし，総評価値が高いものに関連のトピックが複数あるとき，より評価値の高いトピックのみを選択する．

ユーザがダイジェストを視聴している間も，ユーザの視聴履歴からプロファイルを更新する．

3 実験

2006 年 3 月 29 日から 4 月 4 日に NHK 総合テレビで放送されたニュース番組から 373 トピック，データ放送から 580 トピックを収録し，筆者の視聴履歴から作成したプロファイルを用いて 4 月 4 日の 10 分ダイジェストを作成した．プロファイルが異なる 2 ユーザについて，それぞれのダイジェスト結果に選択された上位 3 トピックの総評価値とその内訳 (類似度，放送順序・放送頻度・データ放送のルールによる重要度) を図 2 に示す．ただし，総評価値は類似度と重要度を同じ割合で考慮するものとして計算した．また，各トピックの名前は，結果表示のためだけに手動で与えた文字列である．

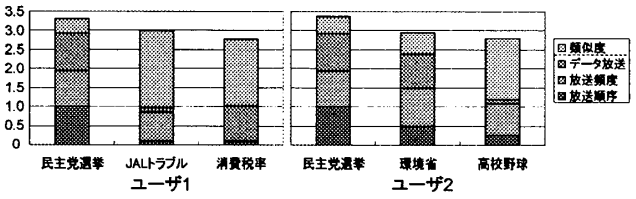

図 2: 実験結果

図 2 より，民主党選挙や JAL に関するトピックのように，類似度と重要度のどちらかが低いトピックでも，もう一方の値が高ければダイジェストに選択されていることがわかる．また，民主党選挙のトピックのように，一日中トップニュースとして報道されているような重要なトピックは，どのような嗜好のユーザであっても視聴できる．図 2 の結果は類似度と重要度を同等に評価したものであるが，それらのバランス (すなわち式 (1) の x, y) を変更することで，異なるダイジェストも作成できる．例えば，視聴者がリラックスして好きな情報を見たいときは類似度を高めに評価し，新しい情報を得たいときは重要度を高めに評価するといったように，視聴者の視聴形態や知的要求に応じたダイジェストを提示できる．

一般に，プロファイルを自動学習する手法は，履歴などを逐次的に反映させていくため，プロファイルが完成するまで時間がかかるという問題があったが，本手法ではシステムの使い始めに重要度を高く評価することで，プロファイルの不完全性を補うことができる．

また，ユーザの操作を監視することでプロファイルを作成する手法では，プロファイルに以前視聴したニュースの情報しか入らないために，同じような内容のトピックがダイジェスト結果となりがちであった．しかし本手法は，類似度だけでなく番組とデータ放送による重要度も総合的に考慮することで，ユーザがまだ見ていないトピックや，新規のトピック，緊急のトピックにも対応できる．例えば，図 2 の消費税率のトピックは，番組では初めて放送されたため放送頻度による重要度は低いが，常時放送され，かつ世情に対して敏感に変化するデータ放送による重要度が高く評価されたため，ダイジェストに選択されている．

4 おわりに

本稿では，視聴者の嗜好とニュースの重要度を総合的に評価してダイジェストを作成する手法について述べた．類似度と重要度の両者を考慮することにより，視聴者の好みだけでなく放送局の意図も反映したダイジェストを作成でき，それらのバランスを変更することで視聴者の視聴形態や知的要求にも適応したダイジェストを提示できる．また，視聴状況や放送状況を自動的に取得して解析することにより，視聴者がプロファイルを記述したり放送局がメタデータを作成するといった作業を削減できる．

今後は，類似度と重要度のバランスや，各ルールの最適なパラメータを自動設定するためのアルゴリズムについて検討する．

参考文献

[1] 松村ほか: “データ放送の視聴者適応提示手法～視聴環境適応型サービス AdapTV の提案とその適応～,” 2005 映情学年大, 19–4 (2005).
[2] 茶筌 Homepage: http://chasen.naist.jp/hiki/ChaSen/.