放送番組に対してパブリックオピニオンメタデータを生成する視聴支援エージェントの開発―ネットコミュニティからの雰囲気成分の抽出とユーザ間での流通による洗練化―

放送番組に対してパブリックオピニオンメタデータを生成する視聴支援エージェントの開発——ネットコミュニティからの雰囲気成分の抽出とユーザ間での流通による洗練化——

岡本　直之[†]　竹之内隆夫[††]　川村　隆浩[†††,†]　大須賀昭彦[†††,†]
前川　守[†]

Viewers-end Agent for Public Opinion Metadata: Generation and Refinement by BBS Community and Viewers Communication

Naoyuki OKAMOTO[†], Takao TAKENOUCHI[††], Takahiro KAWAMURA[†††,†], Akihiko OHSUGA[†††,†], and Mamoru MAEKAWA[†]

あらまし　近年，膨大に蓄積された映像コンテンツを包括的・系統的に整理し，検索性・再利用性を向上させたいというニーズからマルチメディアコンテンツにメタデータを付加することが検討され，記述方法も標準化が進められている．一方，大容量のハードディスクを搭載したディジタルレコーダの急激な普及によって，一視聴者でも大量のコンテンツライブラリを構築することが可能になってきた．このような状況における映像コンテンツの検索には客観的記述をもとにした検索に加えて，他視聴者の反応や盛上りといった主観的な情報を利用した検索も有用になってくると考えられる．本研究ではこの主観的な情報の源として既存のネットコミュニティを利用し，ここで交わされる話題や盛上りをメタデータとして抽出する．これを視聴者側に配置した視聴支援エージェントが行い，更にこのエージェントが視聴者にナビゲーションやユーザインタフェースを提供する．これによりメタデータを利用したナビゲーションに対する視聴者の自然な形でのフィードバックを評価・蓄積・他視聴者で共有することでメタデータを洗練化していく．本論文ではパブリックオピニオンメタデータと呼ぶこれらメタデータの抽出と洗練の手法を説明し，実放送番組に適用して評価した結果を報告する．

キーワード　エージェント，ネットコミュニティ，メタデータ，コンテンツ管理，放送番組

1. まえがき

近年，大容量ハードディスクを搭載したディジタルレコーダが急激に普及している．ディジタルレコーダの登場により，視聴者は従来のテープ媒体の煩わしさから解放され，気ままに番組を録画できるようになった．反面，ディジタルレコーダの普及は新たな要求も生み出した．視聴者は録画の手軽さゆえに次から次へと番組を録画し，視聴する時間が追いつかないため，自分の興味のある，見たいところだけ見たいという要求である．このためには番組の内容を記述したメタデータをコンテンツに付加し，内容による検索を可能にしなければならない．しかし現状，このようなメタデータを番組の映像から自動で付加することは大変難しい．また，メタデータを付加できたとしてもその品質が検索に適したものになっているかという問題がある．

つまり，検索に使用できるメタデータを

- だれがどのように記述し
- いかに精度を向上させるか

という点が問題になっている．本論文では上記の問題に対して，以下のような手法をとる．

（a）　だれがどのように記述するのか

[†] 電気通信大学大学院情報システム学研究科，調布市
The Graduate School of Information Systems, The University of Electro-Communications, Chofu-shi, 182–8585 Japan
[††] 日本電気株式会社，東京都
NEC Corporation, Tokyo, 108–8557 Japan
[†††] （株）東芝 研究開発センター，川崎市
Research & Development Center, Toshiba Corp., Kawasaki-shi, 212–8582 Japan

現在は当該放送番組にかかわる放送事業者またはコンテンツ製作者自身がメタデータを記述することが一般的である．このため，記述される内容は非常に客観的なものになる．再利用性を主眼においてメタデータを付与する場合には当該コンテンツの内容を客観的に表現したものが適切であるが，視聴者にとって価値ある検索を追求した場合にはこれら客観的な表現だけでは必ずしも十分ではない．なぜなら，客観的な記述には内容に対するネガティブな情報，些細な盛上りやある対象へのファン心理といった視聴者の主観的な情報が含まれない．しかし，これらの情報はコンテンツ検索にとって非常に有用である．このため，視聴者が属するコミュニティでどのような受け取られ方をしたか，などという主観的な情報が必要になる．したがって，本研究では放送事業者が付加するメタデータだけではなく，掲示板をはじめとするネットコミュニティでの反応や盛上りの様子を取り込むことにした．

その実現方法として，各視聴者の視聴環境上にソフトウェアエージェントを配置し，このエージェントが視聴者が録画した番組に関係するコミュニティでの反応・盛上りといった主観的データを自動的に収集，メタデータとして検索に利用できる形態に加工するという手法を採用した．

(b) いかに精度を向上させるか

前項ではメタデータを掲示板等のネットコミュニティから抽出すると述べた．しかし，番組を見ながら書き込まれた掲示板の話題が必ずしも番組の内容と一致するとは限らない上，自然言語処理を行うためそのままではどうしても精度が悪いという問題を避けられない．そこで，本論文ではメタデータを使ったシステムを利用するユーザの嗜好やその行動をモニタすることでメタデータの妥当性を判定し，修正する．また，これら修正情報を似た嗜好をもつユーザのエージェント間で随時共有させることで，メタデータの精度を向上させるという手法を取ることとした．

2. パブリックオピニオンメタデータ

パブリックオピニオンメタデータとは，番組が放送された当時の視聴者の生の反応を形式化し，コンテンツに対するメタデータという形で記述したものである．

2.1 視聴者の生の反応とは

視聴者は，放送される内容に対して賛同・批判・歓声・落胆・軽蔑など様々な反応を見せる．こうした視聴者の生の反応を抽出し，時間軸と関連づけて形式化して保存する．この視聴者の生の反応を取得するために本研究では掲示板をはじめとするネットコミュニティを活用することを提案する．[6] をはじめとするこのようなコミュニティでは各局の放送や放送ジャンルに合わせてリアルタイムで放送番組について語り合うチャット・掲示板が存在する．これらは一般に「実況」掲示板と呼ばれる．この情報を利用することで一部のユーザ（「実況」参加者）の生の反応を簡単に得ることができる．本研究ではこれを積極的に利用することとした．

2.2 実況掲示板とは

本節ではこの実況掲示板について簡単に紹介する．実況掲示板は主要放送キー局ごとの各掲示板とジャンルごと，その他掲示板などから構成される（2004 年 8 月現在）．また今年開催されたアテネオリンピックなど大きなイベントの際には臨時に専用の掲示板が開設される．

各掲示板内は任意に作成できる多数（10〜30 程度）のスレッドからなる．スレッドは掲示板内のトピックを表している．内容を見ると中には下品で見るに耐えないものやくだらないものも多数混ざっており，まさに玉石混交である．したがって，このような実況掲示板から有意な情報を抽出するためには，実況掲示板に特化したフィルタリングの技術が必要になる．

2.3 パブリックオピニオンメタデータの種類

本論文ではメタデータを以下の三つに分類する．

2.3.1 一次メタデータ

一次メタデータは放送事業者やコンテンツ製作者によって生成され，視聴者が放送番組を録画再生する際に既に存在する情報である．放送番組オフィシャルサイトにおける情報や，スポーツ番組における新聞や通信社による速報記事などがこれに含まれる．これらの情報はコンテンツに対して客観的な事実を伝えるという性質から，コンテンツ内で何が起こったかについては正確な情報を得ることができる．しかし，視聴者の反応や盛上りといったものは検出できず，視聴者が検索を行うシステムのためのデータとしては不十分である．

2.3.2 二次メタデータ

二次メタデータはリアルタイムで放送番組を見た視聴者が構成するコミュニティで交わされた内容から抽出される情報である．これには番組のどこで視聴者が盛り上がったか，どのような話題が交わされていたかといった情報，また視聴者のファン心理や一次情報で

は含まれることのない放送番組あるいは当事者にとってネガティブな情報も含まれる．こうした情報は再利用性を目的としたメタデータには適さないが，視聴者による検索を主眼としたシステムにとっては非常に有用なデータとなる．

2.3.3 三次メタデータ

三次メタデータは二次データを利用したアプリケーションを利用したユーザによるフィードバック情報を加味したメタデータである．フィードバック情報の内容はアプリケーションに依存するが，確実な情報とはいえない二次メタデータをアプリケーションのユーザが補完・洗練させるための情報である．これを二次メタデータに反映することによって洗練化したデータを三次メタデータと呼ぶ．三次メタデータを生成する具体的な例を次項で述べる．

2.4 パブリックオピニオンメタデータの利用例

パブリックオピニオンメタデータを利用したシステムのイメージを図 1 に示す．このシステムでは，パブリックオピニオンメタデータを視聴者側のエージェントが取得し，あらかじめ登録してあるユーザの嗜好情報と照らし合わせ，ユーザに特化した番組要約を実現する様子を表している．ホームエージェントはまず二次メタデータまでの情報から，視聴者の嗜好に特化した要約を提供する．視聴者はこの提示された要約を視聴しながら再生・停止・早送りといった操作を行う．この操作の情報，例えば「ある嗜好 A の属性をもつ視聴者が，ある部分 X を提示されたがその部分をスキップした」という情報から「ある部分 X は嗜好 A とは適合しないのではないか」などという推測が成り立つ．フィードバック情報はこのような情報の集合であり，二次メタデータを補完・洗練させるものである．

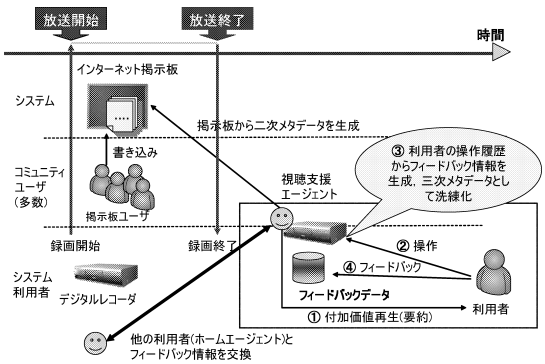

図 1 パブリックオピニオンメタデータの利用例
Fig. 1 An example with public opinion metadata.

このフィードバック情報と二次メタデータを合成していくことでユーザの嗜好に沿った洗練されたメタデータである三次メタデータを得ることができる．

2.5 メタデータの抽出過程

ここではメタデータ抽出の過程として以下のレベル分けを行っている．

2.5.1 Level.1 一次メタデータの収集

最初のステップとして，通信社や番組オフィシャルページから一次メタデータを収集する．ここではプロ野球中継の例を述べる．プロ野球の開催時にはニュースサイトでライブ実況と呼ばれるサービスが提供されている．これは数十秒ごとに Web ページを更新し，現在の試合状況をリアルタイムに提供していくものである．この情報をトレースし，コンテンツの放送時間との同期をとることによって，コンテンツの時間軸上に放送コンテンツ内で何が起きたかを自動的に記述していくことができる．

こうした一次メタデータは，通信社や番組オフィシャルページから提供されるという性質上，コンテンツで何が起こっているかを客観的に記述しただけのものであり，主観的な情報は一切含んでいない．

2.5.2 Level.2 二次メタデータの抽出

このステップでは前述したようにネットコミュニティを利用して二次メタデータの抽出を行う．時々刻々と投稿されるコミュニティ上のメッセージの内容を解析し，雰囲気や話題内容を特徴づける語群を抽出する．この語群はあらかじめ設定した n 個の話題軸からなる n 次元空間上のベクトルとして定義してあり，その出現頻度に応じた形で強度を設定する．これを話題ベクトルと呼ぶ．話題軸は任意の数を設定することができるが，話題軸間の独立性が損なわれないよう注意する必要がある．この話題空間に時間軸を加えた $n+1$ 次元空間に話題ベクトルの時系列的変化を記録していくことで場の盛上りや話題の推移を把握することができる．

またオントロジーを用いて話題要素間の関係をあらかじめ定義し，その強度の比例方法や補強といった手法を加えることでより柔軟で精度の良い話題抽出が可能になるとも予想している．

2.5.3 Level.3 ユーザフィードバック情報によるメタデータの洗練と流通

以下では図 1 で紹介した番組要約システムを例として取り上げ，ユーザフィードバック情報について述べる．このフィードバックデータによって洗練されたメ

タデータが三次メタデータである．

前段階までででユーザに提供するための情報を収集してきたが，この段階ではユーザによるフィードバックを考慮する．まず，ユーザは自分の嗜好を初期情報として事前に設定する．この嗜好情報はキーワードをファイルに列挙する形で各ユーザのローカル環境に保存される．

フィードバック情報とは，アプリケーションによって提示された結果をユーザがどのように受け入れるかを観測することで得られる評価である．この評価から提示した要約結果の妥当性を推測し，二次メタデータを補正する材料とする．

番組要約システムは「実況」から抽出された雰囲気や話題のデータ（二次メタデータ）を参考にユーザの嗜好と類似した話題が交わされている部分を抽出し，ユーザに提示する．このため，これを視聴する際にユーザが行う行動——例えば結果の再生継続・中止など——からその要約結果に対するユーザの満足度を測ることができる．

本研究ではこれらを実現するフレームワークを構築した．このフレームワークはフィードバック情報を生成・利用し，三次メタデータを構成するための統一的な枠組みを提供する．図 2 に概要図を示す．このフレームワークは PC 上で動作し，以後は PC をコンテンツの再生機器として説明を行う．フレームワークは以下の要素から構成される．

- ユーザインタフェース
- 操作履歴データベース
- ルール処理エンジン
- 話題ベクトル補正機構

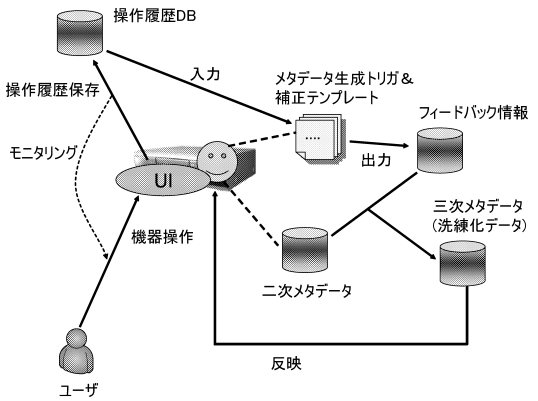

図 2　フィードバック情報生成フレームワーク
Fig. 2　Framework for generating feedback information.

ユーザインタフェースはユーザに対して一般的な映像再生機器と同様のインタフェースを提供し，ユーザは自由に早送りや一時停止などといった操作を行うことができる．ユーザインタフェースはこれらユーザの操作をモニタリングし，操作履歴データベースに蓄えていく．ルール処理エンジンは，蓄えられた操作履歴を逐次読み出し，メタデータ生成トリガと照らし合わせる．メタデータ生成トリガは，フィードバック情報を生成する起因となるユーザの行動を定義したものである．例えば，「提示されたシーンをスキップした」や「シーンを n 秒巻き戻してから最後まで見た」などが挙げられる．メタデータ生成トリガにはメタデータ補正情報が関連づけられている．メタデータ補正情報はメタデータをどのように補正するかを定義したもので，補正対象としては時系列の話題ベクトル値とシーン区切り位置がある．話題ベクトル値に対する補正の場合はトリガに対応するイベントが発生した時間を中心とする時系列ベクトルの形で表現され，シーンチェンジ位置に対する補正ではその補正量が定義される．先ほどのトリガに対応する補正情報としては「当該シーン内においてユーザの嗜好に関係する話題ベクトルを 1/2 にする」や「当該シーンの開始時間を n 秒前倒しする」といった例が挙げられる．話題ベクトル補正機構はユーザの操作から導かれた補正情報（フィードバック情報）を解釈して，エージェントが現在保持しているメタデータと合成することで洗練化されたメタデータ（三次メタデータ）の生成を行う．

なお，このフレームワーク自体は二次メタデータを利用するアプリケーション一般に適用可能であるが，メタデータ生成トリガと補正情報はアプリケーションごとにカスタマイズする必要がある．

ここでプロ野球中継の番組を例に，フィードバック情報から三次メタデータがどのように生成されるか以下に述べる．まず事前にユーザがどの球団のファンか，どの選手が好きかといった嗜好を設定しておく．ここでは巨人軍のファンであると設定したと仮定する．そして要約アプリケーションに対し，自分の好みのシーンを要約して提示するように求める．するとアプリケーションは二次メタデータの情報をもとにユーザの嗜好に沿った，つまり巨人軍の話題が多くかつ興奮・歓喜していた部分という基準でシーンを選択してユーザに提示する．この提示に対し，ユーザはそのまま見たり，スキップしてしまったり，提示されたシーンより少し前にスキップしてから見たり，といった行動を

とる．例えば提示されたシーン A，B に対し，シーン A は即座にスキップされ，シーン B は 15 秒前へスキップしてから最後まで見たとする．これはユーザが暗黙的にシーン A は自分の意に沿わないシーンであり，シーン B について自分の見たいシーンは 15 秒前からであったことを表明していると考えることができる．これにより，シーン A では実際には巨人に関係するシーンではなかったとことが推測され，シーン B では 15 秒前からシーンを提示することが適切であるということが分かる．この推測を具現化したものがフィードバック情報である．これを現在保持しているメタデータと合成することで洗練された三次メタデータが生成される．

更に，このユーザフィードバック情報をユーザエージェント間で流通させる．各エージェントは初期状態として掲示板から生成される二次メタデータをもっている．各ユーザのエージェントはユーザがもつ嗜好と類似の嗜好をもつユーザ同士でクラスタを形成し，その中でそれぞれがもつフィードバック情報を交換する．嗜好の類似性はキーワードが単語単位で部分一致するか，[5] を用いたシソーラス検索において同義語と判定されることを基準とした．異なる嗜好をもつ他のユーザとは交換を行わないため，最終的にフィードバックを反映した結果である三次メタデータにおける話題ベクトルの情報が均質化されてしまうことは多くの場合，回避できると予想される．こうして各エージェントのもつ話題ベクトル情報はユーザの嗜好によって個別に洗練化されていく．

フィードバック情報の流通例を図 3 に示す．

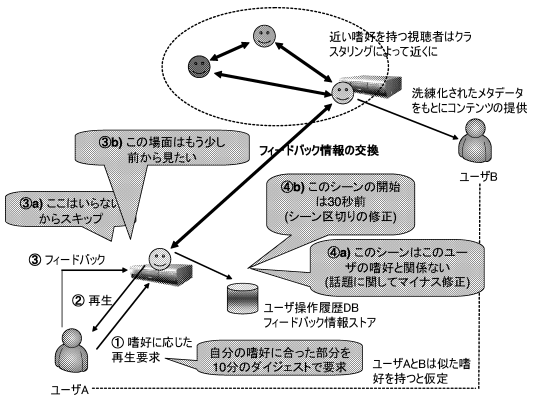

図 3　フィードバック情報の流通
Fig. 3　Circulation of feedback information.

3.　視聴支援エージェントの開発

本論文では掲示板から情報を収集して二次メタデータを構築し，視聴者の番組視聴を支援するエージェントを開発した．本章ではこの詳細について述べる．

3.1　設　　計

メタデータ収集エージェントは以下に述べる六つのモジュールからなる．

- 実況テキスト収集部
- テキスト処理部
- 雰囲気・話題抽出部
- エージェント間通信部
- コンテンツ要約部
- 洗練化ロジック部

3.1.1　実況テキスト収集部

本モジュールは，外部のサーバに随時書き込まれる「実況」の発言データを定期的に収集し，本システムのメタデータサーバ内のデータベースに蓄えていく．

3.1.2　テキスト処理部

従来，テキスト処理の研究では新聞や論文といった文章として整っているテキストを対象にして形態素解析を行うことを前提としているものが多かった．しかし，本研究で対象とするテキストは日本語として全く形になっていない文章も数多く含んでいる．したがって，そのままの形で形態素解析することは非効率であり，このような乱れた言語をどのように正規化していくかが重要になる．

こういったコミュニティでは，コミュニティ独自の文化が形成されていることが少なくない．本研究ではこのようなコミュニティの代表格である [6] を対象としている．このため，このコミュニティに特有の特徴から以下のアプローチによって正規化を試みた．他のコミュニティに適用する場合，必要に応じて各アプローチを取捨選択して組み合わせる必要があるだろう．

（a）　アスキーアートの除去
（b）　同義異字語の正規化
（c）　同音異字語の正規化
（d）　未知語の検出
（e）　外見上類似した同義異字語の補正
（f）　特徴的言換え語の検出

以下では，これら各項目について概要を述べる．

（a）　アスキーアートの除去

アスキーアートとはテキスト文字（狭義には ASCII 文字）のみで構成された文字絵のことである．これら

はテキスト情報としてはほとんど意味がないと考えられるため除去する必要がある．日本で主に用いられるアスキーアートでは一般的には用いられない文字（第二水準文字など）が使われている．したがって，これらの文字を検出することで発言がアスキーアートを含むかどうかを判断できる．この手法は [7] でも用いられ効果を挙げている．本システムでもこの手法を取り入れることとした．

（b） 同義異字語の正規化

同一の意味を表す語でも，そのテキストとしての表現方法は多岐にわたる．仮名文字をとってみても俗に半角仮名と呼ばれる JIS X 0201 片仮名と JIS X 0208 に含まれるいわゆる全角仮名があり，このような掲示板では故意に交ぜ書きされることがある．このような表現の揺れを吸収して解析の精度を向上させた研究として [8] などが挙げられる．また，本研究で新しく対応するものとして「糸吉言侖（結論）」などのような併せ文字が挙げられる．このような文字はワープロ全盛の時代に横 2 倍角文字として強調の意味で多用されていた経緯があり，現代でもチャットのようなくだけた場では擬似的に隠語として使われる．こういった文字を処理するためには一般的な文字に正規化する必要があり，本システムでは併せ文字の表記をデータベース化して置換を行っている．

（c） 同音異字語の正規化

本システムが対象とする掲示板では投稿禁止文字のフィルタにかかることを回避する目的や婉曲的表現，遊び心などから珍妙な当て字が多く見られる．これらの大部分は音読みするともとの語句と同じ読みになるように作られている．本システムではこの性質を利用して，投稿メッセージをいったん仮名読みに分解し，読み仮名から辞書を用いて再変換を試みることで当て字が含まれたメッセージをできる限り本来の漢字使いに戻している．

（d） 未知語の検出

テキスト解析において，辞書に存在しない未知語の存在は大きな問題である．一般的な文章ではその比率も比較的小さいが，本システムで扱うテキストには大きな比率で含まれている．これを検出できなければ後段に控える形態素解析での精度に大きな影響を与えてしまう．しかし，ここで扱う未知語は一般的な文章と違って顔文字をはじめとする感情表現文字など多数の記号類が含まれている．このため従来のように文節を目安に切り分ける手法が通用しない．このため，本論文では以下の手法を考案した．

本論文で扱うテキストには以下の特徴がある．

- メッセージの区切りは明示的に存在する
- 一つのメッセージの長さは短い（1〜3 行程度）
- メッセージ投稿者の同一性が確認できる

以上のことから，各メッセージごとに文字列をパターンとみなして最長に一致する部分を検出し，一致した部分もメッセージとみなし，再帰的にパターンマッチングを行う．このような処理により文字列の一致を関係とした部分文字列の木が生成できる．これを適切に設定したしきい値により短い文字列を除去すると，メッセージ間で共通に使用されている部分文字列が検出できる．この文字列を後段の形態素解析時のためのヒントとすることで精度を高めることができる．

（e） 外見上類似した同義異字語の補正

本システムが抽出対象とするような掲示板では，故意や婉曲的意味を込めてわざわざ似た外見をもつ文字によって語を構成する文字を置き換えることがある．英文字であれば，「i や I ⇔ 1（数字）⇔ l（小文字の L）」「o や O ⇔ 0（数字）」，仮名文字であれば「れ⇔わ」「ぬ⇔め」「ク⇔ワ」「シ⇔ツ」「ソ⇔ン」などが挙げられる．このような文字が出現した場合に，字面そのものだけでなく，考えられる置換え文字を考慮して辞書マッチングを行い，最も可能性の高いものを採用する機構を作成した．

（f） 特徴的言換え語の検出

このコミュニティ [6] では故意に特定の規則に従った語の置換えが見られる．例えば，「促音の長音変化」が挙げられる．これは語の中に促音「っ」があるときに長音「ー」に変化させ後続の文字と位置を交換する．このように規則化できる語の変形がいくつかあり，これを認識することで単語の識別率を向上させることができる．

3.1.3 雰囲気・話題抽出部

雰囲気・話題抽出部では，あらかじめ設定した雰囲気や話題に大きく影響を与える単語の出現頻度を時間間隔ごとに計算し，各単語に与えられた単位話題ベクトルに乗じた後，すべてのベクトルを加算して当該時間間隔の話題ベクトルとした．実際の結果を図 6 に示す．

3.1.4 エージェント間通信部

エージェント間通信部は三次メタデータの流通に際し，P2P ネットワークのピアの役割を果たす．この P2P ネットワークは JXTA に準拠し，ユーザの嗜好

をもとにクラスタを構成してメタデータの効率的な流通を支える．

3.1.5 コンテンツ要約部

コンテンツ要約部は，エージェントが保持しているメタデータ（二次メタデータと三次メタデータ）をもとにして，ユーザの嗜好と合致する部分を抽出してつなぎ合わせ，コンテンツの要約を実現する．具体的には以下の手順でコンテンツを要約して提示する．

（1） コンテンツを画像解析することによってシーンチェンジを検出し，シーンごとに分割する

（2） ユーザの嗜好に合致する話題軸における話題ベクトルの値が事前に設定したしきい値を超える部分を検索し，当該部分が含まれるシーンを選択する

（3） 選択されたシーンを時系列順に結合する

3.1.6 洗練化ロジック部

洗練化ロジック部は，ユーザのインタフェースと連動してユーザがどのような場面を視聴時にどのような操作を行ったかを逐一記録する．これらユーザの操作履歴からフィードバック情報を生成し，三次メタデータを生成する．この内容の詳細は **2.5.3** を参照してほしい．

4. 結果と評価

この章では本抽出手法を実際の番組と関連するネットコミュニティに適用した結果を示し，その正確性・妥当性・有用性を評価する．例示するための題材としてある日のプロ野球中継・巨人–中日戦を取り上げ，実際に起こった事象と抽出されたコミュニティの雰囲気・話題がどのように相関をもつか，一次・二次・三次の各メタデータがどのような精度でコンテンツの内容を表現できるかを検証する．

4.1 共通事項

結果は簡単に可視化するために話題軸は 2 本，これに時間軸を加えた三次元グラフで表現する．話題軸はそれぞれ雰囲気（歓喜・期待・興奮から落胆・失望まで）と話題内容（巨人から中日まで）を設定した．この空間上に各メタデータから得られた情報をプロットした．

4.2 公式記録による結果

ここでは一次メタデータとしてニュースサイトが提供する実況中継データを利用した．これは 30 秒〜1 分の頻度で更新され，現在の打者やその結果，得点経過などがリアルタイムに準ずる精度で得られるものである．このデータから得られる各打者の結果をある時間帯（20:30〜21:00）について，一定の規則で前述の話題空間上にプロットしたものを図 4 に示す．

これを見る限り，当該時間帯には話題になるような事象が存在しないように思われる．

4.3 コミュニティの雰囲気による結果

前項で取り上げた時間帯と同じ 20:30〜21:00 において，二次メタデータによるプロット結果を図 5 に示す．

これを見ると，20:50 ごろに大きな盛上りが見られる．この時間は巨人の投手がリリーフ投手に交代した時間であった．この原因を考察すると，二次メタデータの収集元であるネットコミュニティの性質が浮かび上がる．このネットコミュニティでは投手が炎上する（打ち込まれる）ことに期待する風潮がある．この時間に交代したリリーフ投手はこの試合の前の数試合で何度もリリーフで登板しては打ち込まれることを繰り返していたため，コミュニティとして再度打ち込まれるのではないか，という期待が現れたと解釈することができる．

このようなコミュニティ特有の事情によるベクトル変化はだれにでも普遍的に有用であるとは限らないが，

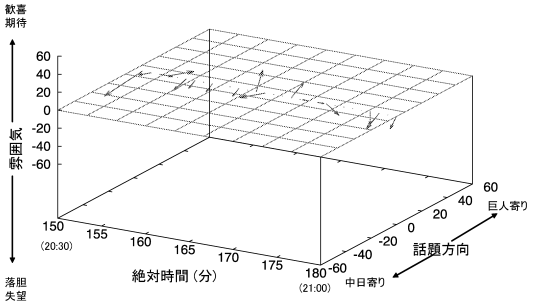

図 4 一次メタデータによる話題ベクトル変化
Fig. 4 Subject vector by primary metadata.

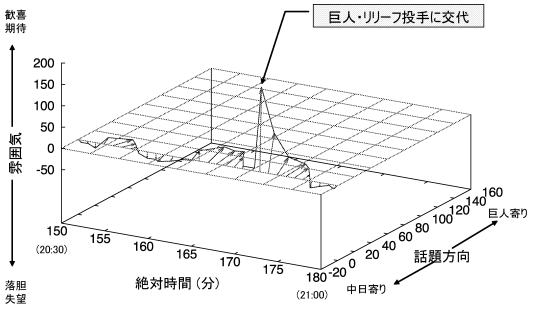

図 5 二次メタデータによる話題ベクトル変化
Fig. 5 Subject vector by secondary metadata.

視聴者がコミュニティの参加者と同等の意識をもっていると仮定すれば非常に有用である．また，二次メタデータの考慮によって，一次メタデータでは検出し得ない場の盛上りを検出できており，二次メタデータの有効性を確認できた．

4.4 誤検出・時間誤差の考慮

二次メタデータのより所は放送番組と同時進行する掲示板のやり取りである．しかし掲示板の参加者が必ずしもその番組の当該時間に起きていることを話題にしているとは限らない．また，番組について書込みを行う場合でも，書込みの発端となった事象とその発言はタイムラグがあるのは当然である．この時間誤差は参加者がすべて同じ反応時間を示すことはないため，常に一定ではなくばらつきがある．この状況を例を図 6 に示す．

この図では 18:45 付近に巨人側に張り出たベクトル群，18:50 付近に巨人軍の選手が放ったホームランによる盛上りが記録されている．しかし，18:45 付近では CM が流れており，巨人に関係する話題が展開するとは考えられない．また，18:50 付近の盛上りの時間には既にホームランを放った選手はホームベースを回りきっており，野球場内の歓声とリプレイが流されている場面であった．このような誤検出や時間的な誤差を低減するために，三次メタデータによるメタデータの洗練機構を提案した．

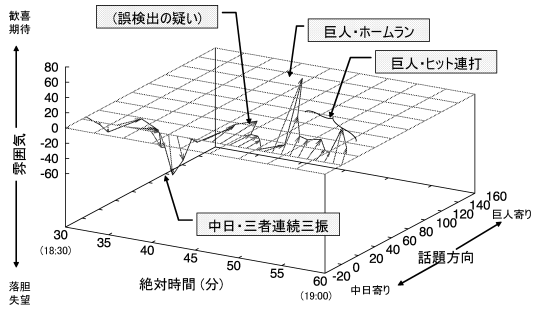

図 6 誤検出と時間誤差（18:30〜19:00）
Fig. 6 False detection and time error.

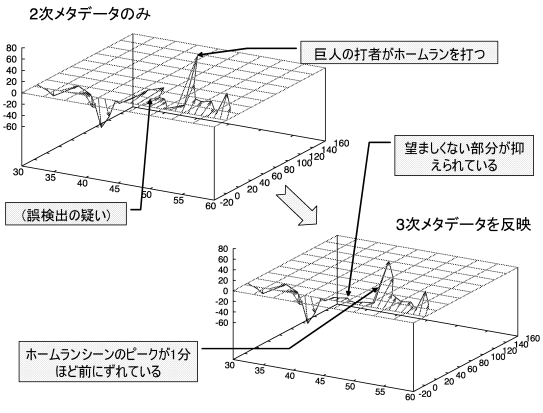

図 7 フィードバック情報による精度向上（三次メタデータ）
Fig. 7 Accuracy improvement by feedback information.

4.5 三次メタデータの反映による精度向上

三次メタデータは二次メタデータにフィードバック情報による補正を加味したメタデータである．この補正の効果を評価するため，5 人のユーザに協力を仰ぎ，もとの二次メタデータに対して検索をして評価を行ってもらった．ユーザの操作履歴からフィードバック情報を生成して，これらを二次メタデータに反映した結果を図 7 に示す．ただし，5 人の被験者はすべて同一の嗜好をもっていると仮定した．

前節で問題となっていた誤検出のベクトルが低減されており，ホームランシーンのピークは時間軸前方向に 1 分前後移動していることが分かる．この効果は三次メタデータを生成するもの，すなわちシステムのユーザが増えるだけ大きくなり，最終的には正確な値に収束していくことが期待できる．

4.6 考　　察

4.6.1 ネガティブ反応・コミュニティ独自の価値観

コミュニティから雰囲気を抽出する本方式では，一次メタデータには現れないネガティブな情報やコミュニティ独自の価値観による盛上りを検出することができる．一次メタデータの客観的な情報だけではこういったコミュニティ独自の盛上りを推測することはできず，コミュニティから二次メタデータを抽出する大きな利点でもある．

4.6.2 コンテンツとの時系列適合性問題

話題ベクトルは一定の時間間隔ごとにベクトルの向きと大きさを計算し，時間軸上にプロットする．話題ベクトルはコンテンツ上での事象に対する視聴者の反応から生成されるために，実際の事象との間にタイムラグを生じる．コンテンツの内容をトレースしながらこの誤差を計測したところ 30 秒〜1 分 30 秒程度であった．これらの誤差は二次メタデータの性質上不可避であるが，三次メタデータを活用することによって低減でき，ユーザが多くなるほど改善されると考えられる．

4.6.3 ピーク誤検出・見逃し

あらかじめ雰囲気・話題を支配する語を決定してお

く本方式では全く対応外の語は何の影響も及ぼさないとして計算している．このため，場の盛上りを検出できずに見逃すことが考えられる．また，語の出現分布によっては存在しないピークを誤検出することがある．しかし，これも前述した三次メタデータやその相互流通を行うことで軽減していくことが可能である．

5. むすび

本論文ではネットコミュニティから放送番組に対するリアルタイムな視聴者の声を得て，その雰囲気や話題を抽出することを試みた．コンテンツに対して視聴者らがどのような反応を示し，どのように盛り上がったかという二次メタデータを抽出し，これを洗練させる三次メタデータを生成・利用する手法を示した．

しかしながら，現状ではいくつかの課題がある．以下にこれらについて述べる．

（a） 番組要約システムの完成と多人数による評価

パブリックオピニオンメタデータの有用性を確認するには，これを利用したシステムの構築が不可欠である．本論文では簡単なコンテンツ要約システムを作成し，有用性を論じた．しかし，その仕組みは単純なものにとどまっている．これを改良し，アプリケーションとして完成させる必要がある．また，メタデータ洗練化の際に用いた被験者数は 5 人と少ない．より多くの被験者によるフィードバック情報を収集し，メタデータを洗練する機構の評価が必要である．

（b） 抽出する単語の選択とスコアリング

本論文では雰囲気・話題を決定づける語を固定として抽出・統計処理を行ったが，自動的に話題の中心となる語を抽出できるような仕組みがあれば便利である．また，話題ベクトル強度の計算方法も改善の余地があると思われる．今後，更にコミュニティを研究して適切な手法を検討していく予定である．

謝辞 本研究を遂行するにあたり，研究の機会と議論・研鑽の場を提供して頂き，御指導頂いた国立情報学研究所/東京大学本位田真一教授をはじめ，活発な議論と貴重な御意見を頂いた研究グループの皆様に感謝致します．また，研究環境を提供して頂き，異なる視点から有益な御意見を頂いた電気通信大学の小林良岳助手と中山健助手に感謝致します．

文　献

[1] “MPEG-7 Multimedia Description Schemes WD (Version 1.0),” 1SO/IEC JTC1/SC29/WG11 N3113, Maui, Dec. 1999.

[2] 浜口斉周，道家　守，林　正樹，“MPEG-7 メタデータを用いた自動番組制作システムの検討,” 第 65 回情報処理学会全国大会予稿集 4E-3 3-37, 38, 2003.

[3] 石井　恵，中渡瀬秀一，富田準二，“名詞句と単語の勢いを用いた話題抽出手法の提案,” 情報処理学会研究報告「自然言語処理」，no.2003-NL-160, 2003.

[4] 冨浦洋一，田中省作，日高　達，“共起データに基づく名詞の n 次元空間への配置,” 情報処理学会研究報告「自然言語処理」，no.2002-NL-154, 2002.

[5] “シソーラス（類語）検索,”（株）言語工学研究所，http://www.gengokk.co.jp/thesaurus/

[6] “2 ちゃんねる,” http://www.2ch.net/

[7] 松村真宏，三浦麻子，柴内康文，大澤幸生，石塚　満，“2 ちゃんねるが盛り上がるメカニズムの解明,” 第 51 回人工知能基礎論研究会，SIG-FAI51-7, 2003.

[8] 竹元義美，福島俊一，“口語的表現を含む日本語文の形態素解析の実現と評価,” 情報処理学会研究報告「自然言語処理」，no.1994-NL-103, 1994.

（平成 16 年 11 月 24 日受付，17 年 3 月 25 日再受付）

岡本　直之

2003 電通大・電気通信・情報工卒．2005 同大大学院情報システム学研究科博士前期課程了．同年 4 月，同大学院情報システム学研究科博士後期課程入学，現在に至る．主としてエージェント，メタデータの研究に従事．

竹之内隆夫

2003 電通大・電気通信・情報工卒．2005 同大大学院情報システム学研究科博士前期課程了．同年 4 月，日本電気（株）入社，現在に至る．主としてエージェント，ユビキタスコンピューティングの研究に従事．

川村　隆浩

1992 早大・理工・電気卒．1994 同大大学院理工学研究科電気工学専攻修士課程了．同年（株）東芝入社．2001〜2002 米国カーネギーメロン大学ロボット工学研究所客員研究員．現在（株）東芝研究開発センター知識メディアラボラトリー研究主務，工博．2003 より電気通信大学大学院客員助教授．主としてマルチエージェントシステム，セマンティック Web の研究・開発に従事．情報処理学会，人工知能学会各会員．

大須賀昭彦　（正員）

1981 上智大・理工・数学卒．同年（株）東芝入社．1985〜1989（財）新世代コンピュータ技術開発機構（ICOT）に出向．現在（株）東芝研究開発センター知識メディアラボラトリー研究主幹．工博（早大）．2002 より電気通信大学大学院客員教授並びに大阪大学大学院非常勤講師兼任．主としてソフトウェアのためのフォーマルメソッド，エージェント技術の研究に従事．1986 年度情報処理学会論文賞受賞．情報処理学会，日本ソフトウェア科学会，IEEE CS 各会員．

前川　守

1965 京大・工・数理卒．同年東京芝浦電気（株）入社．東大理学部情報科学科助教授等を経て，1993 電通大大学院情報システム学研究科情報システム設計学専攻教授，現在に至る．Ph. D. 主としてオペレーティングシステム，分散処理，ソフトウェア開発環境，マルチメディアの研究に従事．ACM，IEEE，情報処理学会各会員．