

アノテーションに基づく デジタルコンテンツの高度利用 (後編)

日本アイ・ビー・エム (株) 東京基礎研究所

長尾 確

knagao@jp.ibm.com

マルチメディア・コンテンツの高度利用

本稿の前編(前号)では、主にアノテーションに基づくWebドキュメントの高度利用について述べたが、当然ながら、アノテーションは、マルチメディア・コンテンツに関しても大きな役割を果たす。一般にマルチメディア・コンテンツはテキスト・コンテンツと異なり、キーワードによる検索ができないし、内容の要約や分類も容易ではない。

そこで、人間の手によるメタ情報を使って、好みのビデオなどを効率よく発見する手法などが提案されている。

まず、これまでに行われてきたマルチメディア・コンテンツに関するメタデータの標準化に関する活動を紹介する。

マルチメディア・コンテンツに関するメタデータ

マルチメディア・コンテンツに関するメタデータは、元々テレビ放送用のコンテンツ管理への応用を目指して設計されている。たとえば、アメリカのSMPTE (Society of Motion Picture and Television Engineers) では、放送用コンテンツのプロダクションおよびポストプロダクションにおける制作者のために、主に機械制御での利用を想定したメタデータエレメント、符号化、

素材に関する識別子を基本としたメタデータ体系を規定している。

また、ヨーロッパのEBU (European Broadcasting Union) では、1998年にP/Meta Projectを発足させ、テレビ番組制作者、放送事業者、アーカイブ業者などの間でのメタデータ相互運用のための標準化を進めている。

これらは、放送業界にかかわるプロ向けのメタデータ標準であったのに対し、主に消費者が放送コンテンツを利用するためのメタデータ標準として、次に紹介するMPEG-7、TV Anytimeがある。また、主にコンテンツ管理者のためのメタデータ標準として、後述するコンテンツIDがある。

MPEG-7

MPEG-7 (Moving Picture Experts Group Phase 7) は、ISO/IECに属するMoving Picture Experts Group (MPEG) によって標準化活動が行われている新しい規格である⁴⁾。この標準は、マルチメディア・データの内容を記述するための枠組みを規定し、デジタルライブラリ、マルチメディア検索、番組選択、編集などのアプリケーションの開発や普及に寄与することを目的とする。

MPEG-7は、マルチメディア・データの内容を記述する記述子(画像におけるオブジェクトの色やテクスチャ、動画におけるオブジェクトの動きや位置などのような特徴の表現形式)の集合を規定する。この記述がコンテ

ンツに付与されることにより、マルチメディア・コンテンツの内容に基づく検索が可能になる。XMLに準拠した、記述子 (Descriptor) や記述子間の関係 (Description Scheme. DS) を定義する言語 (Description Definition Language. DDL), また、メタデータとしての記述を伝送・蓄積するシステムの仕様などが規定される。記述する対象には、静止画、グラフィクス、3Dモデル、オーディオ、スピーチ、ビデオ、アニメーションなどの構成要素である。MPEG-7は主にメタデータの仕様であり、参照するコンテンツの符号化法や蓄積法には依存しない。たとえば、アナログの動画や紙に印刷した写真に対する記述を構成することも可能である。

MPEG-7の詳細および現在の状況については、Webサイト (<http://www.cselt.it/mpeg>) を参照していただきたい。

TV Anytime

TV Anytimeは、1999年7月に結成されたTV Anytime Forumによって策定が行われている、テレビ番組コンテンツを検索・選択するためのメタデータとその利用のモデルである⁵⁾。これは、ハードディスクなどのストレージ利用やインターネット利用をベースにした、放送と通信におけるマルチメディア・コンテンツの国際的な相互流通システムの実現を目指している。TV Anytimeのフェーズ1では、次の5つの課題について標準化を議論している。

1. ビジネスモデル

想定するサービスとそのプロファイル、技術仕様を検証するためのベンチマークアプリケーションなどを規定。

2. システム

全体のシステムのアーキテクチャとAPIモデル、システム内の機能モジュール、モジュール間インタフェース、アプリケーション例に関する動作モデルなどを規定。

3. メタデータ

メタデータの利用モデル、ECG (Electronic Content Guide)、コンテンツの検索・ナビゲーション、蓄積制御に関するメタデータ構造とその伝送方式などを規定。

4. コンテンツリファレンス

コンテンツのIDのフォーマットとその参照方式、コンテンツ配信情報であるロケータのフォーマット、放送やインターネットを利用したロケーション解決方法などを規定。

5. ライツマネージメント

ライツマネージメントモデル、コンテンツの著作権管理保護方式、個人認証方式、蓄積課金方式、アクセス制御、コンテンツ著作権管理保護制御用メタデータなどを規定。

TV Anytimeのメタデータは、コンテンツ作成の段階

からセットで伝送され利用されるものと、コンテンツとは別に編集され、特別な用途に対応するために別途伝送され利用されるものがある。メタデータの用途は、検索、ナビゲーション、ロケーション解決、セグメンテーションがある。セグメンテーションに関するメタデータは、番組を冒頭から見のではなく、特定の内容だけを選択的に視聴するためのものである。このデータは、MPEG-7と整合性をとるために、XMLベースで記述される。

メタデータによって好みの番組の検索が終わり、その番組のコンテンツリファレンスID (CRID) が取得できた場合、CRIDから最終的なロケータを取得するために、いくつかのプロセスが想定されている。目的の番組が複数の番組から構成されている場合、あるCRIDから複数のCRIDを得て、それぞれのロケータを取得することになる。また、あるCRIDに対して複数の取得手段が選択できる場合には、複数のロケータが一度に得られる場合もある。最終的に、決定されたロケータに従って、目的のコンテンツをアクセスすることになる。

CRIDの構文は以下ようになる。

CRID://<authority>/<data>

<authority>は、RFC1591に準拠したドメイン名で、<data>は、RFC2396のURIに準拠したリソースIDである。たとえば、CNN (CNN.com) が所有するHeadlineNewsというタイトルの番組は、CRID://CNN.com/HeadlineNewsと記述できる。

TV Anytimeは、MPEG-7との整合性のため、メタデータ構造記述にMPEG-7 DDLを使用している。ただし、MPEG-7にないDSを新規に規定している。たとえば、Media Review DSと呼ばれる、評論家の解説情報がある。これと、MPEG-7のUser Preference DSを組み合わせ、ユーザの好みのコンテンツを自動選択するアプリケーションが考えられる。

コンテンツID

コンテンツIDは、2000年10月に設立されたコンテンツIDフォーラムによって規定されている、主に著作権管理を目的としたメタデータ標準である⁶⁾。それは、コンテンツを識別するユニークコード、コンテンツ制作時に定まる権利属性、コンテンツ流通時に定まる権利運用属性、流通属性、ロイヤリティなどの分配属性を含んでいる。これらの属性の他に、流通業者などがアプリケーションやメディアに応じて属性を拡張することを可能にする自由領域、コンテンツIDの検証の目的で追加するシステム管理情報も含まれる。

コンテンツIDが、MPEG-7やTV Anytimeのメタデータと大きく異なる点は、ユニークコードが電子透かしによって、コンテンツに埋め込まれ、切り離せない状

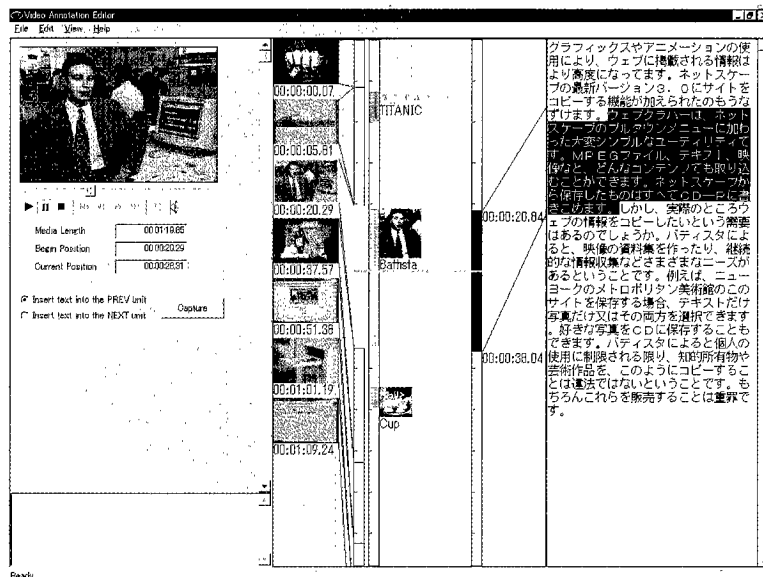


図-1 ビデオアノテーションエディタ

態になっているということである。ただし、電子透かしは常に利用されるとは限らず、ユニークコードは RDF (Resource Description Framework) に準拠した DCD (Distributed Content Descriptor) と呼ばれるメタデータ記述形式で表現され、コンテンツにバインドして流通させることもある。

コンテンツ流通後に変更の可能性がある属性については、コンテンツID 管理センタと呼ばれるサーバの知的財産権管理データベースで集中的に維持管理され、ユニークコードをキーとしてアクセスすることができる。ただし、このデータベースは、アクセス制御がかけられており、属性項目ごとにアクセス権が設定されている。

マルチメディア・アノテーションとビデオ・トランスコーディング

マルチメディア・コンテンツのメタデータあるいはアノテーションは、これまで主に著作権管理、検索、選択などの応用を目指して設計されてきているが、さらに多くの応用を可能にする枠組みが、本稿の前編でも紹介したセマンティック・トランスコーディングである⁸⁾。

映像コンテンツをトランスコーディングする場合、まず映像コンテンツに含まれる音声のトランスクリプト(書き起こしたテキスト文)を用意する。このトランスクリプトに、意味構造や、シーンの変わり目のタイムコード、シーンごとのキー・フレームの位置、映像の各シーンに登場するオブジェクトの名前とその出現位置(時間と座標)などをアノテーションとして付加する。

セマンティック・トランスコーディングシステムでは、トランスクリプトを自動的に生成して、半自動的にアノテーションを作成できる。映像のシーンの変わり目も自動認識し、シーンに関するタグ付けを支援する。

このシステムは、現在のところ、映像コンテンツの要約の生成、映像コンテンツからテキストと画像からなるコンテンツへの再構成や、ビデオ音声の翻訳などが実現できる。

図-1はビデオアノテーションエディタの画面例を示している。このエディタは、ビデオのシーンへの分割と音声部分のテキスト化を行う。自動処理の結果はインタラクティブに修正できる。図-2はビデオのアノテーションに基づいて作成した要約ビデオを再生するプレイヤーの画面例である。要約ビデオモードでは、要約部のみを再生し、フルビデオモードでは、任意のシーンをランダムに選択・再生できる。

映像を要約するには、まず映像のトランスクリプトを要約する。その要約に対応する映像シーンを抽出することによって映像の要約を実現している。映像シーンの抽出は、タイムコードの情報を手がかりに自動的に行う。映像コンテンツからテキストとイメージへの変換は、クライアント側にビデオ再生機能がない場合に有効に使えるだろう。映像コンテンツ中に含まれる、それぞれのシーンを代表する画像とそれぞれのシーンの内容を表すテキスト文からなるコンテンツを生成することができる。さらに、生成したテキスト文を要約/翻訳することも可能である。近い将来に映像コンテンツの音声部分を翻訳し、映像と同期させながら合成音声で出力する機能も統合する予定である。それによって、1つの映像コンテンツから複数の言語に対応した映像コンテンツを作成することが実現できるだろう。

筆者は映像コンテンツが今後重要な情報ソースにな

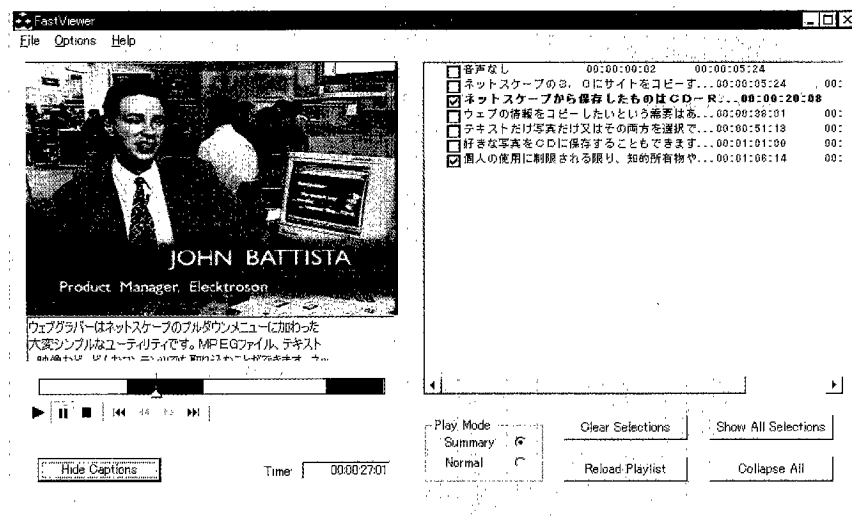


図-2 要約ビデオプレイヤー

ることを確信している。そのため、要約やフィルタリングに限定されない、コンテンツの再利用を可能にするさまざまな枠組みをできるだけ早めに用意しておきたいと考えている。ここでのアノテーションを利用した手法は、将来の枠組みに対しても容易に付加情報を変換して対応できるようにした。たとえば、前述のMPEG-7のようなアノテーションの標準的なフレームワークが確立した場合にも容易に移行できる。

現在、映像コンテンツの要約はテキストの要約と同様に盛んに研究が進められている。古くはCarnegie Mellon Universityが開発したInfomediaがある¹⁰⁾。これは、映像コンテンツに含まれるさまざまな属性を自動抽出して、より重要な部分を選択する。たとえば、画面上に現れる文字情報や人の顔、シーンの変わり目、クローズド・キャプションと呼ばれる字幕情報などを使う。あらかじめリストアップされた重要な固有名詞の出現頻度や、キーワードの重要度を計算し、そのキーワードの現れるシーンをつなぎ合わせて要約とする。

他の例としては、IBM Almaden Research Centerが開発しているCueVideo¹¹⁾や、同T. J. Watson Research Centerが開発しているVideoZoom⁹⁾が挙げられる。

CueVideoは映像コンテンツ中のキー・フレームを並べて表示し、人間がキー・フレームのどれかを選択し、その部分の映像のみを再生することによって、映像コンテンツ全体を見る手間を減らすことができる。またCueVideoでは、紙芝居のように静止画を表示しながら音声再生する手法も採用した。これは、シーンの変化時だけ静止画を入れ替え、コンテンツのダウンロードに要する時間を節約することをねらったものである。音声は再生スピードを変化させることによって、早口にしたたり、ゆっくり聞き取りやすくすることもできる。この他、音声認識を利用した映像シーンの検索も実現

されており、任意の単語やフレーズを入力すると、音声認識でその言葉を含む部分を抽出してリストアップすることができる。

VideoZoomでは、映像の解像度をシーンに応じて動的に変化させる。たとえば、解像度の低い映像ををまずダウンロードして、細かく見たいところのみについて差分の情報を追加していくことができる。この手法も、ネットワークや表示デバイスの制約に依存して、映像コンテンツを加工するトランスコーディングの一種といえる。

トランスコーディングの仕組み

セマンティック・トランスコーディングの具体的なシステム構成としては、トランスコーディング実行用のソフトウェアを、プロキシサーバ側に置いた。プロキシサーバは、ユーザがPCなどのクライアント側からの要求に応じて所望の結果を返す。さらにアノテーション情報や事例を収めたアノテーションサーバを別個に用意した。こうした形をとるのは、不特定多数の人にソフトウェアを利用してもらうことで、サーバ側の事例辞書にノウハウを蓄積し、自動処理の精度を高めるためである。

セマンティック・トランスコーディングを実行する複数のソフトウェア・モジュール(トランスコード)は、HTTP(HyperText Transfer Protocol)プロキシ上で機能するプラグインとして実装した。トランスコードを制御するHTTPプロキシをトランスコーディングプロキシと呼ぶ。

図-3はセマンティック・トランスコーディングシステ

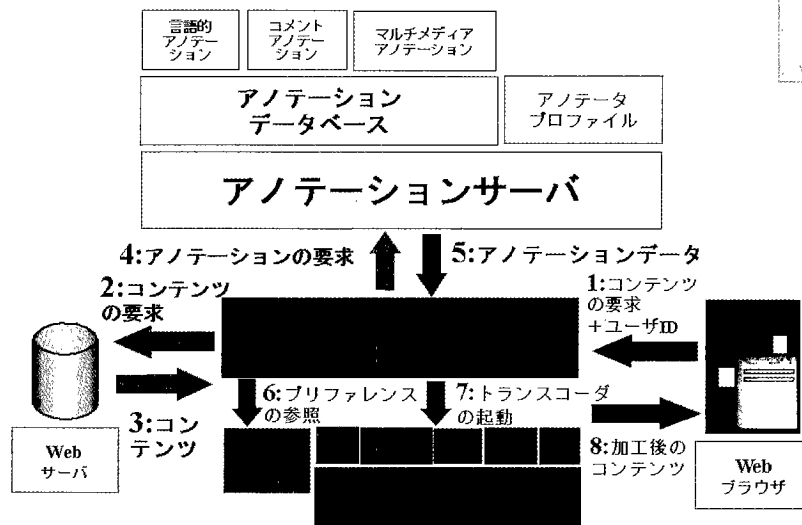


図-3 セマンティック・トランスコーディングシステムの構成

ムの構成を表している。

トランスコーディングプロキシを中心とした情報の流れは次のようになる。

1. クライアントのWebブラウザからURL (Uniform Resource Locator) とクライアントIDを受け取る。
2. WebサーバにURLの示すWebページをリクエストする。
3. Webページを受け取ると、そのハッシュ値を計算する。
4. アノテーションサーバにURLに関連するアノテーションデータを要求する。もし、アノテーションデータが見つかったら、アノテーションサーバからデータを受け取る。
5. データを受け取ると、データのハッシュ値とWebページのハッシュ値と比較する。
6. 同時に、クライアントIDに基づいてユーザ情報を検索する。ユーザ情報がない場合は、ユーザから与えられるまでデフォルト設定を使う。
7. ハッシュ値を照合したら、アノテーションデータとユーザ情報に基づいて適切なトランスコーダを起動する。
8. 加工したコンテンツをユーザのWebブラウザに送信する。

トランスコーディングプロキシは、実装環境としてIBM Almaden Research Centerの開発したWBI (Web Intermediaries) を使用した^{☆1, 3)}。このWBIを利用したトランスコーディングプロキシには、以下の3つの主要な機能がある。個人情報の管理、アノテーションデータの収集と管理、そしてトランスコーダの起動と結果の統合である。

個人情報の管理を行うには、まずアクセスしてきたユーザを特定する必要がある。ユーザの特定にCookieを使う。個人情報を管理するIDを、Cookieデータとし

てユーザに渡す。これにより、ユーザのアクセスポイントに関係なくユーザの特定が行える。ただし、既存のWebブラウザは、Cookieをセットしたサーバに対して、そのCookieを渡すものであり、プロキシのCookie利用は考慮されていない。通常プロキシは、ホスト名とIP (Internet Protocol) アドレスのみによってユーザを識別する。そこで、ユーザが個人情報をセットした時に、Cookie情報 (ユーザID) と個人情報を関連付け、一方、アクセスポイントの変化ごとにIPアドレスとホスト名、Cookie情報 (ユーザID) を関連付け直す。これによりIPアドレスが変化してもユーザの特定が行える^{☆2}。

トランスコーディングプロキシは、アノテーションサーバと通信して、アノテーションデータを入手する。アノテーションサーバは複数存在することができるので、それぞれのサーバの管理するアノテーションデータのインデックスを定期的に作っておく。このインデックスを、どのアノテーションサーバからデータ入手すべきかを判断するときに役立てる。トランスコーディングプロキシの最も重要な役割は、個人情報とアノテーションデータに基づいてコンテンツを加工することである。コンテンツの加工は、必要なトランスコーダを起動し、その結果を統合することによって行う。現在、開発済みのトランスコーダは、テキスト文、画像、音声、映像にそれぞれ対応したものである。これ

☆1 WBIは、IBMのWebサイトであるalphaWorks (<http://www.alphaworks.ibm.com/>) からダウンロードできる。WBIは、プログラマブルなHTTPプロキシであり、通常のプロキシとしての機能の他に、ユーザごとのアクセス制御や、プロキシに流れるデータの加工を容易に行えるAPI (Application Programming Interface) を提供する。

☆2 通常のプロキシとして動くときは、クライアントID (ホスト名とIPアドレス) → Cookie情報 (ユーザID) → 個人情報という流れで、クライアントIDから個人情報を引き出す。アクセスポイントが変化したときは、プロキシをWebサーバとしてアクセスすることで、Cookie情報を取得し、クライアントIDとCookie情報 (ユーザID) を関連付け直す。

らのトランスコーダは、直列あるいは並列に結合することで、複合的なトランスコーディングが実現できる。たとえば、文書を要約後に翻訳して、さらに音声化するなどの一連の処理をトランスコーダの使い分けにより行う。

展望：アノテーションとトランスコーディングによってもたらされるもの

アノテーションは、従来のデジタルコンテンツを知的コンテンツとするための最良の手段である。それは、人間が、自分自身あるいは他者の創り出したコンテンツを再評価し、価値あるものとそうでないものを見分けるよい機会が与えられるからである。コンテンツを人類共有の財産とするためには、やはりそのコンテンツを責任を持って吟味する人間が必要であろう。アノテーションとは、まさにそのような責任の所在を明らかにし、内容にさらなる価値を与えていく仕組みなのである。

また、トランスコーディングは、コンテンツのアクセシビリティ（ユーザの身体的特性やスキル、使用するツールなどによらずに適切にアクセスできること）を強化する手段である²⁾。これによって、コンテンツは真に人類共有の資源となる。

昨今鳴り物入りで登場したSemantic Web⁷⁾が機械のためのWebなら、アノテーションとトランスコーディングによって拡張されたWebは、人間と機械がよりよく助け合って利用するためのWebである。システムの内部に人間が上手にかかわっていくための仕組みがないと、知識共有のような高度なシステムはうまく機能しないだろう。

アノテーションは人間と機械の構成するシステム全体が賢くなっていくための仕組みである。この場合の機械とは、あらかじめプログラムされた手続きを文脈に応じて選択的に実行する自律的なシステム、すなわちエージェントである。エージェントをある程度以上に複雑にする代わりに、コンテンツの方をアノテーションによって、人間がエージェントにとって都合のよい形に変えていけば、人間とエージェントとコンテンツが構成するシステム全体をより高度にすることができる。つまり、コンテンツそのものがより理解しやすくなれば、それを扱うエージェントが可能なタスクもより高度になるだろう。エージェントはアノテーションの付与されたコンテンツを対象にすることによって、単純な手続きを繰り返すだけで、より高度なサービスを提供できる。これは、見かけ上、エージェントが賢くなったように見えるが、実際はコンテンツそのものが（人間の不断の努力によって）賢

くなっているのである。

このようになって初めて、人々はエージェントの価値を認めて受け入れていくだろう。そして、情報の収集や分類などのタスクはエージェントに任せて、より創造的な仕事に専念できるようになると思う。

おわりに：情報の洪水を乗りきるために

今後の課題には、もちろん、アノテーションの作成コストを下げていくことが含まれるが、その他に、アノテーションに基づく、Webコンテンツからの知識発見を実現することがある。近い将来には、Web上の情報検索には、既存の検索エンジンではなく、複数のコンテンツから新たな知識を得てその結果を要約して出力するような、いわば知識発見エンジンを使うようになるだろう。それによって、ハイパーリンクを集めた大量のリストの代わりに、短時間で容易に理解できるように要約されたコンテンツを読むことができるようになると思われる。さらにもう1つの課題は、映像や音声といったマルチメディア・データを含むデジタルコンテンツの効率的な検索である。この場合の検索の質問には単なるキーワードではなく、音声あるいはテキストの自然言語文を用いるだろう。

こうした課題を克服することは、将来やってくる情報の洪水から自分自身を守る最良の方法になるだろう。オンライン・コンテンツを人類共有の知識とするために一丸となって努力をすることがなければ、人々は今後も無限に拡大していく情報の圧迫から自分自身を解放することができないだろう。

参考文献

- 1) Amir, A., Srinivasan, S., Ponceleon, D. and Petkovic, D.: CueVideo: Automated Indexing of Video for Searching and Browsing, In Proceedings of SIGIR '99 (1999).
- 2) Asakawa, C. and Takagi, H.: Annotation-based Transcoding for Nonvisual Web Access, ACM ASSETS 2000 (2000).
- 3) IBM Almaden Research Center: Web Intermediaries (WBI), <http://www.almaden.ibm.com/cs/wbi/>
- 4) 柴田正啓: MPEG-7の標準化動向, 映像情報メディア学会誌, Vol.55, No.3, pp.337-343 (2001).
- 5) 栗岡辰弥: TV Anytime Forumにおける標準化動向, 映像情報メディア学会誌, Vol.55, No.3, pp.344-352 (2001).
- 6) 阪本秀樹: Content ID Forumの標準化動向, 映像情報メディア学会誌, Vol.55, No.3, pp.353-358 (2001).
- 7) 浦本直彦: Semantic Web - 機械のためのWeb -, 人工知能学会誌, Vol.16, No.3, pp.412-419 (2001).
- 8) Nagao, K., Shirai, Y. and Squire, K.: Semantic Annotation and Transcoding: Making Web Content More Accessible, IEEE Multimedia, Vol.8, No.2, pp.69-81 (2001).
- 9) Smith, J. R.: VideoZoom: Spatio-temporal Video Browser, IEEE Trans. Multimedia, Vol.1, No.2, pp.157-171 (1999).
- 10) Smith, M. A. and Kanade, T.: Video Skimming for Quick Browsing based on Audio and Image Characterization, Technical Report CMU-CS-95-186, School of Computer Science, Carnegie Mellon University (1995).

(平成13年1月7日受付)