アノテーションに基づく意味的メディア統合*

長尾 確
名古屋大学 情報メディア教育センター
nagao@nuie.nagoya-u.ac.jp

1 はじめに

メディア情報を意味的に統合するためには、メディア情報の日常世界へのグラウンディング、つまり情報と人間の世界との明示的な関連付けが必要になる。情報のグラウンディングに対するアプローチには、メディア情報の内容を人間の直感に適合する形で機械的に操作可能な記号表現で記述する知識表現 (あるいは、オントロジー) アプローチと、メディア情報に対する機械的な処理を人間が補足して内容記述を作成していくアノテーション・アプローチがある。本研究は、アノテーション・アプローチによってメディア情報のグラウンディング、さらに意味的統合を目指すものである。具体的には、テキストやマルチメディアコンテンツに対して半自動的にアノテーションを作成して、検索や要約や翻訳、さらに複数のコンテンツ間の関連付けを行う。アノテーションには、音声のトランスクリプトや自然言語文の言語構造、映像のシーンやオブジェクトに関する意味属性 (カテゴリー情報) などが含まれる。

2 セマンティック・アノテーションとトランスコーディング：意味的に共有可能なデジタルコンテンツ

ここでは、人間と機械の間で意味的に共有可能なデジタルコンテンツを目指したアプローチについて述べる。そのようなコンテンツを作り出すためには、「情報のグラウンディング」つまり、機械が処理する情報の、人間の世界との明確な関連付けを行わなければならない。これは、人間と機械が情報の意味を共有するために不可欠なプロセスである。

近年、Web 上でこのようなグラウンディングを扱うためのプロジェクトが W3C において提案されている。いわゆるセマンティック・ウェブ [14] がそれである。セマンティック・ウェブの最初の基本的な目標は、コンテンツに十分に定義された形式的記述を関連付けることである。この記述は、Web アプリケーションの開発者が、ある特定のコンテンツからそれに関するさまざまな属性を抽出し、その配信において考慮すべき問題を明確にするために利用される。

セマンティック・ウェブは現在の Web と切り離して設計されているのではなく、その情報の意味を明確に定義された Web の拡張版として設計されている。近い将来に、機械は現在は表示のみに注意が向けられている情報の意味を処理でき、それによって、機械と人間はよりよく協調できるだろうということである。

セマンティック・ウェブは主に、コンテンツに含まれている概念の形式的な記述を要求するという意味でオントロジーに基づくアプローチを採用していると言える。これは、このような形式的な記述が一つのコンテンツに留まらず、より普遍的なものであるべきから、トップダウンなアプローチである。

もう一つのアプローチ、つまりアノテーションに基づくアプローチは、ここでの最も重要なトピックである。アノテーションは、ある特定のコンテンツに関わるものであり、この修正はインクリメンタルに行えることから、ボトムアップなアプローチと言える。当然、アノテーションを持つコンテンツは、機械的な理解や処理を促進し、高い精度で個人に適合したコンテンツに加工することができる。これは、やはり情報のグラウンディングを実現する一つの手段である。

オントロジーに基づくアプローチとアノテーションに基づくアプローチは、矛盾するわけではなく、お互いに補完的な関係にある。具体的なコンテンツがすでにある場合は、意味的な内容記述はコンテンツを密接に関連付けることができ、そのようなコンテンツがない場合は、より一般的な概念記述を詳細に定義し、具体的なコンテンツとの連結の仕方を定義しておく必要がある。前者は、アノテーションに基づくアプローチであり、後者はオントロジーに基づくアプローチである。これらをどのように統合するかは、さらなる研究が必要である。

*Annotation-Based Semantic Media Integration by Katashi Nagao (Center for Information Media Studies, Nagoya University)

以下では、セマンティック・ウェブにおけるコンテンツの意味とオントロジーに関してより詳しく議論する。

2.1 コンテンツのセマンティックスと情報グラウンディング

セマンティック・ウェブが機能するためには、機械は構造化された情報とその情報を使って推論をするための推論規則にアクセスする必要がある。人工知能の研究者は、Web が発明されるはるか以前から同様のシステムに関する研究を行ってきた。「知識表現」と呼ばれるこの技術は、まったく正しいアイディアであったにも関わらず、世界を変えるほどのことはまだ起こっていない。この技術によって実現可能な重要なアプリケーションは無数にあるが、そのすべての機能を実現するには、ある単一のグローバルなシステムにその技術を整合的に統合する必要があるのである。

セマンティック・ウェブの研究者は、それとは異なり、多用途性を達成するためには不整合さを許容し、すべての質問に答えることができないことを始めから認めるという立場をとっている。彼らは、Web において十分な範囲で推論ができる程度の記述力を持つ推論規則用の言語を開発している。けっして全域的において整合である仕組みを作ろうとしているのではない。これは、初期の Web において、中心的なデータベースや明確な構造を持たずに、組織化されたデジタルライブラリができるはずがなく、ゆえにすべての必要なものを発見することなど不可能である、という批判がなされたことに似ている。このような批判は正しかったが、システムの記述力は広範囲の情報を利用可能にすることに成功した。また、10 年前には不可能だと思われていた、Web コンテンツの十分な規模のインデックスも作成可能になってきている。セマンティック・ウェブのチャレンジは、コンテンツと推論規則の両方を記述する言語を設計することにより、これまで知識表現のために構築されてきたさまざまな技術を Web 上で利用可能にすることである。

2.1.1 オントロジー

セマンティック・ウェブでは、情報の意味は情報資源の意味記述フレームワークである RDF [12] で表現される。RDF は XML 形式による 3 つ組で構成され、その基本構文は主語、動詞、目的語に相当する。主語と目的語は情報資源の識別子である URI (Universal Resource Identifier) [11] によるリンクで指示される対象であり、動詞は対象間の属性 (A は B の一種とか A は B の一部とか) を表現する。ただし、動詞も Web で誰かが定義しているものを使う場合は、同様に URI で参照することができる。

もちろん、これで話は終わりではない。なぜなら、Web 上に 2 つのデータベースがあってそれぞれが同一の概念に異なる名前をつけていたとしたら、機械はこれらが同じ意味であるかどうかを判別しなければならなくなるからである。理想的には、機械はどのようなデータベースに遭遇しようとも、共通の意味を持つ名前を自動的に発見できる仕組みを持つことが望まれる。

この問題に対処するために、セマンティック・ウェブでは、3 つめの基本的コンポーネントである「オントロジー」と呼ばれる情報の集合を規定する。哲学では、オントロジーは存在論と呼ばれ、世界に存在する対象が持つ存在の本質に関わる体系のことを指す。人工知能研究者は、オントロジーを語彙の間の関係を形式的に記述するデータとして扱ってきた。オントロジーの最も典型的な例は、語彙間の階層構造や語彙間の推論規則の集合である。

語彙の階層構造は、記述対象のクラスやそれらの関係を定義する。たとえば、住所はある場所のタイプ、町コードなどで表現されるが、それらは同一の対象を指していると考えられる。複数の対象の間の関係をクラスやサブクラスによって記述することは Web においても重要なツールとなる。対象間の無数の関係は、クラスに属性を割り当て、サブクラスがクラスの持つ属性を継承することで表現される。もし町コードが町を表すあるクラスで規定されていて、町がその Web サイトを持っているとき、町コードとその Web サイトの間の直接的な関係がデータベースに定義されていなくても、一般的な町クラスの持つ町コードと Web サイトの関係を参照することで、容易に町コードからその Web サイトの URI を調べることができる。

オントロジーにおける推論規則はさらなる力を提供する。たとえば、「ある町コードがある国コードに関連していて、ある住所がある町コードを使っていれば、その住所はその町コードに関連した国コードにも関連する」という規則が記述できる。それによって、あるプログラムは、名古屋大学は名古屋にあり、名古屋は日本にあるので、名古屋大学は日本にあり、日本の住所フォーマットに準拠して住所を記述すべき、ということが容易に推論できる。コンピュータがそのような情報を正しく「理解」しているとは言えないが、ある種の語彙操作を効果的に行い、人間にとって意味のあるものを自動的に生成することは可能になる。

Web にオントロジーを載せることで、語彙のミスマッ

チに関する問題に対処する明確な方法ができつつある。Web コンテンツに用いられる言葉の意味は、そのコンテンツからオントロジーの項目へのポインターによって定義される。このような問題はずっと以前から存在した。人の名前のことを、氏名や姓名などと表現して、異なるデータのように記述してもそれらの指す内容が同じであるといったことは頻繁に起こっている。これはオントロジーにおいて同値関係として定義することで解決できる。つまり、氏名と姓名と名前が同じ意味であるという関係を定義して、オントロジーに含めるのである。

クラス間の意味的な同値性が定義できれば、一方の形式から他方への自動的な変換が可能になるだろう。たとえば、あるサービスは、ユーザーの「名前」を要求しており、あるデータベースにはそのユーザーの「氏名」が記述されているとき、オントロジーの定義を参照することで、それらの同値性を認識し、自動変換によって「氏名」データを「名前」データに変換してサービスに渡すことができる。これによって、サービス間のミスマッチを解消して、より高度な相互運用性が実現できるだろう。

このようにオントロジーは、Web の機能を拡張することができると思われる。たとえば、Web の検索は、オントロジーを利用することで、曖昧なキーワードを用いる場合と異なり、特定の概念に関連のあるコンテンツのみを選択することができるため、ひじょうに精度の良いものになるだろう。より高度な応用は、Web コンテンツに含まれる情報をある知識の構造や推論規則に関連付けることであろう。

2.1.2 情報のグラウンディング

人間社会において情報技術を十分に有益なものにするためには、デジタルデータの日常世界に対する「情報グラウンディング」を行う必要があるだろう。それは、[?] で述べられているように、機械と人間の間で情報の意味を共有することである。

いわゆるシンボルグラウンディング問題 (記号接地問題とも呼ばれる) [2] とは、もともと、デジタル情報は、人間の解釈によって外部から与えられるのではなく、それ自身が初めから持っているべき実世界での意味を、どうしたら持ちうるのだろうか、という問いであった。

もともとの議論におけるグラウンディングは、実世界というのは物理世界のこととして扱われる傾向が強く、それゆえ主にロボティックスの研究者の関心を集めていたが、現在の議論においては、実世界とは人間の日常世界のことであり、それは、物理的な側面のみならず、概念的、社会的、その他の側面を合わせ持っている。

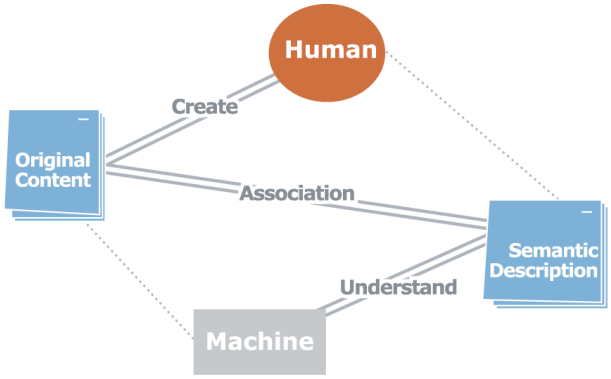


図 1: コンテンツと意味記述の間の関連性と情報グラウンディング

このようなグラウンディングによって、われわれは「知的コンテンツ」と呼ばれるものを手に入れることができる。オリジナルのコンテンツは、自然言語による文書や映像・音声などさまざまなモダリティに関わり、それらの情報グラウンディングによって、より精度の高い検索、翻訳、要約、可視化などが可能になる。知的コンテンツを推進する活動には、前述のセマンティック・ウェブの他に GDA (Global Document Annotation) [3] や MPEG-7 [5] などがある。

情報のグラウンディングによって、デジタルコンテンツを日常世界とリンクし、知的コンテンツを目指すアプローチには、主に以下の 2 つがある。

1. 知識表現あるいはオントロジーに基づくアプローチ
2. アノテーションに基づくアプローチ

これは、図 1 にあるように、相互補完的な関係にあると考えられる。

3 セマンティック・ウェブとセマンティック・トランスコーディング

次世代の高度な知識共有のインフラであるセマンティック・ウェブセマンティック・ウェブに不可欠なものは、ユーザーが自由にコンテンツを作成し、さらにその共有化を促進し、知識として再利用可能にするためにコンテンツを意味的に拡張するツールやプラットフォームである。われわれは、これまでセマンティック・トランスコーディングという枠組みにおいて、現在の Web の若干の延長線上に、セマンティック・ウェブと同等の機能を有する仕組みを研究してきた。それは、主に以下の 3 つのシステムから成っている。

1. 既存の Web コンテンツに対して、機械による内容理解を促進する補足情報 (アノテーション) を付与するツール
2. アノテーションデータを管理･共有するためのサーバ
3. アノテーションデータを用いてコンテンツを動的にユーザに適合させるためのトランスコーディングプロキシ

これらのシステムを実現した経験に基づいて、セマンティック・ウェブの早期実現のために、近い将来に、われわれが何を、どのように行うべきか、に関する一つの提言を行いたいと思う。

セマンティック・ウェブ [14] はグローバルな知識共有のインフラを目指すアプローチであるが、そのような試みがそう簡単にうまくいくはずがないのは、これまでの人工知能 (特に、知識工学)、あるいは経営工学における知識管理などの分野における方法論の多くの失敗から考えても明らかである。ではどうしたら、この困難な問題に対処できるだろうか。

筆者の考える「現在の」セマンティック・ウェブの主な問題点は以下の通りである。

1. RDF や OWL 等のセマンティック・ウェブを構成する記述言語の仕様が必要以上に複雑になっている。
2. メタデータの作成の方法論が確立していないため、何をどう作ればよいのかわからない。
3. 意味的な検索以外の具体的な応用についてほとんど述べられていない。
4. マルチメディアデータのようなテキスト以外のコンテンツの意味構造化についてはまったく述べられていない。

以下では、これらの問題についての具体的なアプローチについて述べる。キーとなる概念は、アノテーションによる情報の階層化とトランスコーディングによる情報の個別化である。当然ながら、これらは XML のような、情報を相互運用するための枠組みを前提としている。

アノテーションとトランスコーディングによって、内容および利用の観点から拡張された Web は、「現在における」セマンティック・ウェブの一つの具体例と考えられる。これは、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関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、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良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 [1]。これによって、コンテンツは真に人類共有の資源となる。

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良い形に変えていければ、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。

4 アノテーション：デジタルコンテンツの階層化

図 2 で示されるように、アノテーションは、現在の Web に上位構造を作る基盤になる。現在の Web コンテンツが最下層で、アノテーションはコンテンツに情報を付け加えるメタ (上位) コンテンツ、さらにメタコンテンツに対するメタコンテンツのように階層をなしている。

この図にあるように、従来の Web コンテンツは一枚の平面上に存在する要素群として捉えることができる。

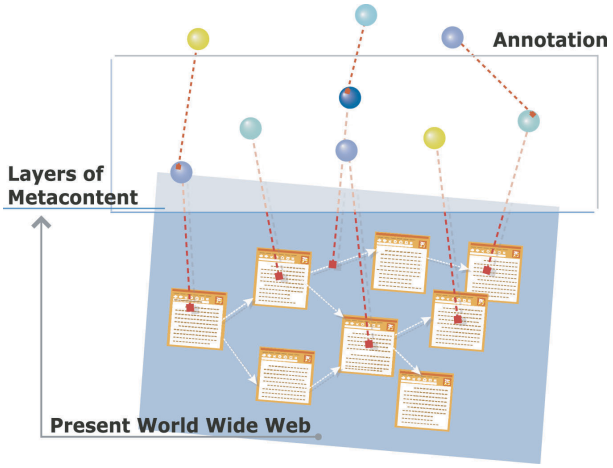


図 2: アノテーションによる Web コンテンツの拡張

セマンティック・トランスコーディングでは、Web コンテンツを平面から立体に拡張する手法を提案する。コンテンツの各要素に意味や文書構造を示すアノテーションを付加する。このことによって Web コンテンツに、コンテンツの各要素の意味や文書構造を記述した上位構造を築くことができる。代表的なアノテーションの例としては、リンク元の文書に埋め込まれていないハイパーリンクである外部リンク XLink[10] や、コンテンツに対するコメントなどが挙げられる。アノテーションを作成して公開することが容易になれば、Web コンテンツの表現力は大幅に高まり、その利用価値が飛躍的に向上するだろう。

アノテーションによる階層化の手法を用いて、具体的には HTML 文書などの Web コンテンツが抱える、以下の 3 つの課題を解消できるだろう。

1. HTML ではレイアウトなどの文書の表現については規定している。しかし、文書の意味などといった内容に関してはほとんど何も規定していない。この点を改善するために、RDF[12] を用いることができるが、その方法論は未だ確立しているとは言えない。

2. HTML などで記述したハイパーテキストは、各文書間のネットワーク構造を記述できる。ただしリンク情報が常に正しいとは限らず、その修正ができるのはもとの文書の著者だけである。

3. Web 文書の著者は一般にその読者のことを考慮して著作してはいない。なおかつ著者と読者の間に立って吟味・調整する役割の人間も通常はいない。

Web は、新しいスタイルの文書のあり方を示したという点において革新的だったと言えるだろう。Web コンテンツの自由度の高さは疑いようがない。しかし、現状では Web コンテンツを読者が読みやすいような体裁に機械的に変換することは非常に困難である。

5 トランスコーディング：デジタルコンテンツの個別化

デジタルコンテンツがあたりまえのものとして世の中に溢れ出したのは 20 世紀の情報技術の進歩からすると必然的であっただろう。そして、それら膨大なコンテンツを活用するための技術もさまざまなものが発明され、進歩を遂げていくことは間違いがない。これまでは、ともかくコンテンツを作成して流通させることが主目的であったのに対し、これからは、それらのコンテンツをいかに賢く利用するか、あるいは、いかに多様に、多目的に利用するか、ということが最も重要な課題になると思われる。

デジタルコンテンツの高度利用の主なものに、パーソナライゼーションとアダプテーションがある。デジタル放送の映像や Web ページなどのデジタルコンテンツをユーザの好みに応じて変換することをパーソナライゼーションと呼び、それらのコンテンツを PC や PDA や携帯電話などのデバイスの特性に合わせて変換することをアダプテーションと呼ぶ。これらは、ともにコンテンツの個別化の例である。個別化はコンテンツの送受信がブロードキャストからポイント to ポイントになったことと大いに関係がある。

ここでは、デジタルコンテンツのパーソナライゼーションとアダプテーションを合わせたものをトランスコーディングと呼ぶ。現状では、オンラインコンテンツへのアクセスは PC 経由で行なわれることが多い。しかし、この様相は近年、急激に変わりつつある。PC に加えて、携帯電話や PDA、テレビ、カーナビなどを使ってインターネットにアクセスする機会がますます増加するだろう。このとき重要となるものがトランスコーディングである。たとえば、PC で表示することを前提にして作成した Web ページを携帯電話などで表示する場合、画像の縮小やテキスト部分の圧縮といった操作を自動的に行なう必要がある。トランスコーディングには、少ない伝送容量を使ってサーバからクライアントにコンテンツを配信できるという利点の他に、ユーザの嗜好に応じた理解しやすいコンテンツを生成できるといった利点がある。トランスコーディング技術を使えば、画面の表示機能やデータ伝送速度など、それぞれ違った仕様や制約をもつ多様な機器に対して、1 つのコンテンツ・ソースから情報やサービスを提供できるようになる。コンテンツ・プロバイダやサービス・プロバイダは、それ

それの機器に対応したコンテンツを個別に用意しなくても済む。具体的な応用例としては、PC 向け Web コンテンツのトランスコーディングによって、携帯電話向けのコンテンツを生成するといった利用法がある。コンテンツ・プロバイダは、現状のように PC 向けと携帯電話向けのコンテンツを作り分ける必要がなくなる。

このトランスコーディングをさらに進めて、テキストの要約などの内容に基づく処理の精度を高める工夫を盛り込んだのが、筆者の提案するセマンティック・トランスコーディングである [7]。具体的には、コンテンツに含まれるテキスト文要素に言語構造や語彙情報をアノテーションとして関連付けることによって、要約や翻訳などの自然言語処理の精度を大きく向上させることができる。たとえば、アノテーションによってコンテンツに含まれるテキスト文の意味を明確にすると、正確な要約や翻訳が期待できる。コンテンツにアノテーションを付ける手間が増すが、誤解なく伝達すべき重要な情報にはアノテーションを付与して、より適切な形で伝達・共有すべきだろう。このアノテーションはコンテンツの内容理解を促進するものとして機能する。

セマンティック・トランスコーディングは、ユーザが指定した Web 上の新聞記事などのコンテンツを任意の圧縮率で要約して表示したり、テレビ番組などの映像データからユーザの好みに応じた話題だけを抜き出して、ダイジェスト映像を作成するといったことを可能にする。さらに、要約したコンテンツを翻訳したり、テキストを音声化して聴くこともできる。

コンテンツサーバにおかれたテキスト、画像、音声、映像などのコンテンツはトランスコーディングプロキシによって、ユーザの使用するデバイス (PC、携帯電話、カーナビなど) や、ユーザの要求 (概要をつかみたい、母国語で読みたい、声で聞きたい、など) に合わせて加工される。このとき、アノテーションと呼ばれる付加情報を用いて、より精度の高い要約・翻訳を行う。アノテーションはアノテーションサーバに蓄えられている。

セマンティック・トランスコーディングは、基本的にテキストコンテンツの処理を中心としたものであるが、その手法は映像や画像などの非テキストコンテンツの加工にも応用され、マルチメディア・データを含むコンテンツに適用できる。

6　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 (トランスコーダ) は、HTTP プロキシ上で機能するプラグインとして実装した。トランスコーダを制御する 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 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 [4]。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なうには、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 アドレスのみによってユーザを識別する。そこで、ユー

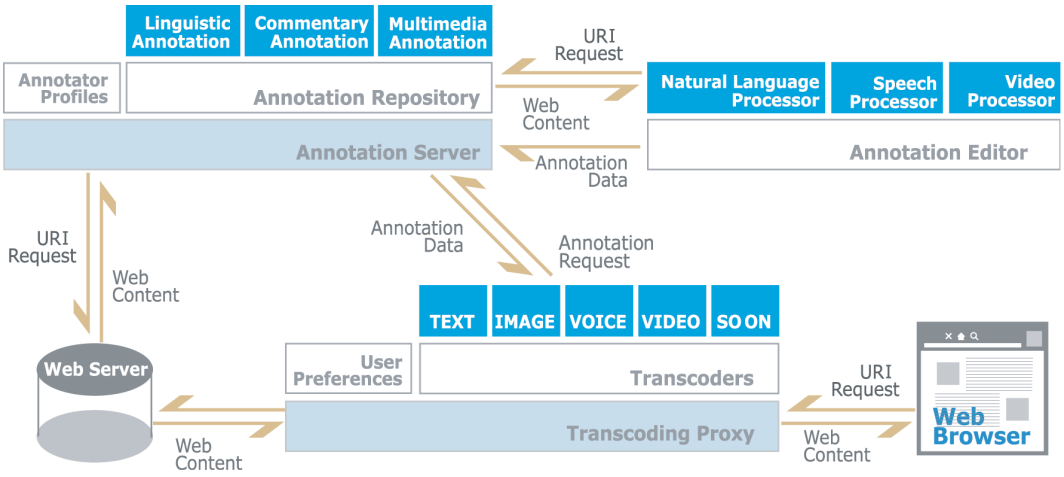


図 3: セマンティック・トランスコーディングシステムの構成

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。

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を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なう。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。

7 セマンティック・ウェブへの提言

セマンティック・トランスコーディングで用いるアノテーションは、主に文書の言語構造、マルチメディアの内容に基づく構造化情報、任意のコンテンツに対するコメント情報などである。これらは、ある種のリテラシーがあれば誰にでも作成可能な情報である。そのようなリテラシーは、アノテーションエディタと呼ばれるツールを使っているうちに自然に獲得されていくような仕組みにすべきだと思っている。

一方、「現在の」セマンティック・ウェブにおける主な (メタ) コンテンツは、RDF によるグラフ構造によるメタデータは何をどう作ればよいのかよくわからないし、OWL[15] によるオントロジカルなデータは、さらに何をどう記述すればよいのかわからない。やはり、具体的なコンテンツに関して、自然に追加できるような内容でないと動機的にも技量的にもとっかかりがないのである。

アノテーションは、コンテンツと乖離したトップダウン的なものであるべきではないし、段階的により高度なものに発展させていく必要があるだろう。そのために必要なのは、コンテンツをサーバ側で変換して配信する場合にも、オリジナルデータへのポインター (データベース URL やレコード ID など) を変換後のコンテンツの該当する部分に挿入し、アノテーションがオリジナルデータに直接関連付けられるように工夫することである。

また、当然ながら、セマンティック・ウェブは、現状の Web とシームレスに統合できるものでなければならない。トランスコーディングプロキシはサーバとクライアントの「中間」で処理を行なうため現在の Web のアーキテクチャに自然に統合される。ここで必要なのは、URL のようなコンテンツのポインターを要求するだけでなく、どのプロキシにどのような変換を必要するかということも含めて要求とすることである。これは、現在ではブラウザの機能とトランスコーディングプロキシのデータベースを用いることで解決しているが、たとえば、トランスコーディングのためのプロファイル

を XML を用いて標準化して、SOAP (Simple Object Access Protocol)[13] 等でリクエストを送るようにすれば、より一般化できるだろう。

8 マルチメディアコンテンツのアノテーションとその応用

セマンティック・ウェブには、具体的なコンテンツの意味構造化に関する視点がまったく欠けていると言わざるを得ない。文書に対する言語構造のアノテーションに関しては別稿に譲ることにして、ここでは、マルチメディアコンテンツのへのアノテーションとその応用について述べる。

映像や音声を含むマルチメディアコンテンツは、テキストコンテンツに比べて、内容に基づく処理が極めて困難である。そこで、マルチメディアコンテンツの検索・変換を行う上で必要となるインデックス情報を生成・加工し、これらをアノテーションとしてオリジナルコンテンツに関連付けておく、という手法が考えられる。われわれは、特にビデオデータを対象に研究を進めている [8, 22, 6]。

ビデオデータがさまざまな視聴環境で利用される状況を考える場合、その対処方法としては以下の 2 通りが考えられる。

- 異なる視聴環境ごとに適合させたデータをあらかじめ用意しておく
- オリジナルデータを環境に合わせてオンデマンドに変換する

セマンティック・トランスコーディングが採用しているのは、後者の方法である。すなわち、コンテンツ提供者はデータを 1 種類のみ用意すればよく、利用者側の環境や要求に応じてサーバがコンテンツ変換を行うようなシステムである。この場合の変換は、画面サイズやビットレートなどの視聴環境のパラメータを用いて行なう単純な変換の場合と、要約や翻訳のようなユーザの好みと (アノテーションによってもたらされる) コンテンツの意味情報を同時に考慮して行なう複雑な変換の場合がある。セマンティック・トランスコーディングは、特にアノテーションに基づくさまざまなコンテンツ加工を特徴としている。

ただし、これを可能にするためには、コンテンツ変換を容易にするアノテーションを適宜付与しておく必要がある。コンテンツ変換を容易にするアノテーションとは、すなわちそのコンテンツの内容記述である。ビデオデータであれば、データ中のシーンやシーンに含まれる対象、人物の発話内容などである。

以下では、われわれが試作しているマルチメディアコンテンツ検索システムの概要について述べ、アノテーションに基づいたビデオ検索を実現する手法を現状のシステムを例に提案、解決すべき問題について考察する [22]。

8.1 先行研究・事例

完全なビデオデータを対象とした研究や実例は少ないが、単語と動画像の相互検索 (ただし、音声ではなくクローズドキャプション (字幕) のような映像に付加されているテキストを利用)[19]、話者と発話内容の同時検索 [21]、Web 上の文書と画像のクロスメディア検索 [18]、画像から画像の検索、ニュース音声のトランスクリプトに対する検索 [20] など、これまでに多くの研究がなされており、岡らのグループが研究・開発した CrossMediator[17] は、現在実用化されているマルチメディア検索システムの 1 つである。彼らは、音素や濃度ヒストグラムといったデータの時系列特徴量をインデックスとして用いて検索を行っている。

音声データを音素系列として記述する方法は、音声認識誤りによる検索精度の劣化や未知語に対して強いという利点を持つ一方で、同音語や単語境界誤り、短い単語による精度劣化や、内容に基づいた検索や要約が難しいという欠点がある。

個人が家庭で録画したデータに対して検索を行う場合には、このような手法が適していると考えられるが、Web 上で無数のデータが公開されるような場合、それらに対して検索・要約を行うためには十分な内容記述が必要である。

8.2 ビデオ検索におけるアノテーションの有用性

8.2.1 テキスト検索・イメージ検索との比較

例として、次のような検索要求を考えてみる。「テロで航空機がビルに激突したらしい」

テキスト検索ではどうだろうか。検索エンジン Google(www.google.com) で And をとる単語を増やしながら検索した結果を表 1 に示す。

表 1: Google によるテキスト検索例

検索キーワード	ヒット件数
テロ	376,000
テロ, ビル	76,500
航空機, ビル	30,700
テロ, 航空機	25,500
テロ, 航空機, ビル	11,400
貿易センタービルに航空機が激突	90

検索キーワードとしてユーザが与える語数は平均 1,2 語と言われているが、テキストコンテンツの量がビデオコンテンツとは比較にならないほど多いことを考慮しても、検索結果としてユーザが確認・視聴するためには、相当の絞り込みや提示方法の工夫が必要であることが予想される。

一方、イメージ検索について考えてみると、この例の場合、「航空機」、「ビル」、「爆発・炎上」といったイメージあるいは単語列から検索を行うことになる。

イメージからの検索は精度面から考えて現状では非常に困難であるため、ここでは単語列からの検索のみについて考える。同様に Google のイメージ検索を利用した結果を表 2 に示す。

表 2: Google によるイメージ検索例

検索キーワード	ヒット件数 (正解)
テロ	1,700 (計測不能)
テロ, ビル	117 (15)
航空機, ビル	44 (6)
テロ, 航空機	37 (6)
テロ, 航空機, ビル	11 (4)

上記の正解件数は、テロに関係するものをカウントしており、この中で実際に航空機が写っているものはわずかであった。これは、イメージの説明文としてニュース記事の文章を利用していることにより、航空機という単語が補完された結果であると考えられる。逆に、ニュースのように公共性の高い情報でなく十分な説明がなされていない場合には、イメージの検索は容易ではない。これはビデオ検索についても同様に当てはまる。

8.2.2 ビデオデータの特徴

ビデオデータの最大の特徴は、映像が持つ過去の事象の偽らざる連続性である。テキストには、それを扱う人間の知識や状況が反映され、時間軸の前後やスキップも容易である。イメージには、瞬間の情報は凝縮されているが、時間情報が欠落している。

ビデオデータをテキスト (実際には音声) とイメージの合成と考えると、補完の最も難しい情報は時間の経過によって我々が得ることのできる真実である。

先ほどのテロの例を考えてみる。同様の例として、イタリア・ミラノで起きた小型機事故と合わせて、テロという事実情報に着目してみる。

米国同時多発テロの場合は、事件の初期段階では、テロという言葉は断定的には使われていない。大統領の声明により、その瞬間から事件はテロであるとの認識が定まったわけである。このため、1 機目と 2 機目の激突のシーンに対してテロという情報を与えるには、その真実を理解した人間の介在が必要になる。

一方、イタリア・ミラノ小型機事故の場合は、事件の第一報で、「テロ攻撃の可能性が非常に高い」との見方が示されたと報じたが、その後に事故説と自殺説が出た。このため、事件直後のテロという情報が誤っていることを与えるためには、やはり同様に人間の介在が必要である。

このように、ビデオデータの場合は、必ずしもその時点での音声が真実を表しているとは限らない。そのため、後からアノテーション情報を付与することが可能な仕組みが非常に重要であると考えられる。

8.3 アノテーションを利用したコンテンツ変換例

アノテーションが付与されているデータは、様々な自然言語処理が可能である。その例として、ビデオデータから HTML ドキュメントへのコンテンツ変換を行った例を図 4 に示す。また、図 5 は、このビデオドキュメントを要約した例を示す。さらに、図 6 と図 7 に、それぞれ PDA 用と携帯電話用に変換されたビデオドキュメント示す。

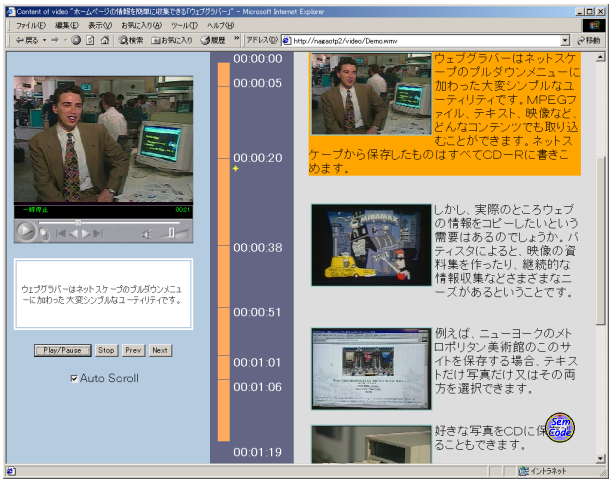


図 4: ビデオドキュメント

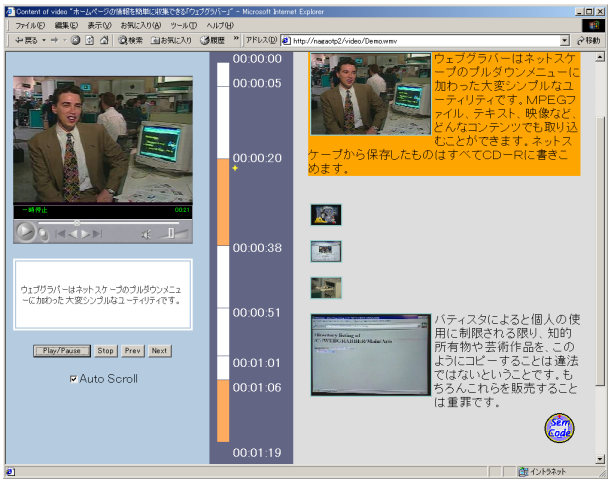


図 5: 要約されたビデオドキュメント

図 6: PDA 用に変換されたビデオドキュメント

図 7: 携帯電話用に変換されたビデオドキュメント

このように、検索されたビデオデータを、ユーザの利用環境や嗜好に応じて適宜変換して提示することも、アノテーションを付与することによって可能になる。一度に多くの検索結果を視聴することのできないビデオデータにとって、これらの処理を可能にするアノテーション情報は必要不可欠である。

9 アノテーションに基づくビデオ検索システム

9.1 システム概要

我々が提案するアノテーションに基づいたビデオ検索システムの概要を図 8 に示す。

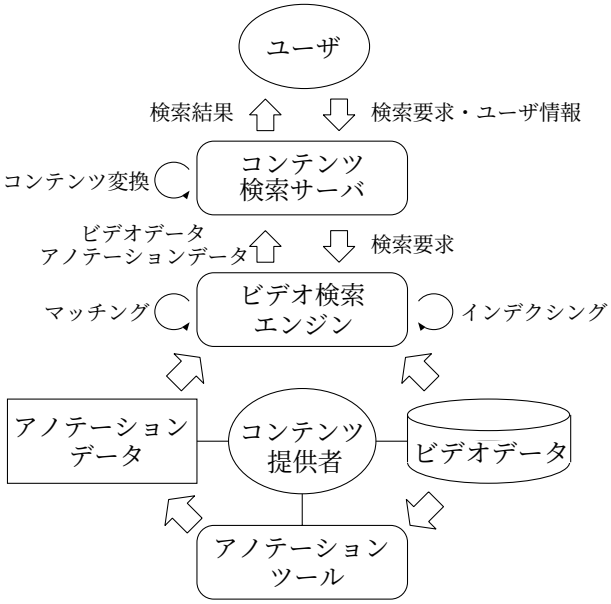


図 8: ビデオ検索システムの概要

システムは、コンテンツ提供者側によるアノテーション生成処理、ユーザからの検索要求に対する検索処理、ユーザ環境に応じた検索結果の変換処理、の大きく 3 つに分けられる。

ビデオ検索画面を図 9 に示す。これは、通常の Web ブラウザを使い、Web ページの検索と同様に、キーワードや自然言語文を入力して、関連するビデオへのリンクを含む検索結果を得ることができる。

検索結果の変換処理については、セマンティック・トランスコーディング [7] によって行う。

図 9: ビデオ検索画面

9.2 アノテーション作成

我々は、ビデオデータ中の発話情報、シーン情報、シーン内オブジェクト情報をアノテーションとして生成・編集・関連付けすることを可能にするツールとして、多言語ビデオトランスクリプタを開発している (図 10)。

ビデオデータへのアノテーションの付与は、図 11 に示される手順にしたがって行われ、作成したアノテーションデータは、XML で記述される。

アノテーションによって扱う内容記述は、シナリオにおけるト書きのような情景描写 (シーン) と登場人物の台詞に相当する発話文 (トランスクリプト)、およびフレーム内に登場するオブジェクトの記述から構成される。

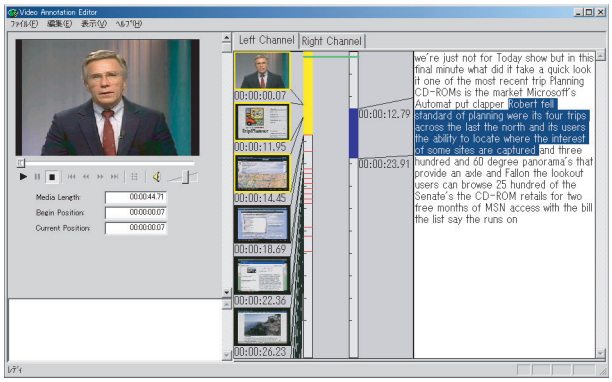
発話情報にはタイムコードと発話内容、シーン情報にはタイムコードとインデックスタイトル、オブジェクト情報にはタイムコードと名称、説明、矩形、リンク情報等が含まれる。

現在、音声認識、言語識別、シーン検出は自動的に処理されているが、認識誤り訂正、シーン統合、オブジェクトの切り出し、シーン記述等は、人間が行う仕組みになっている。

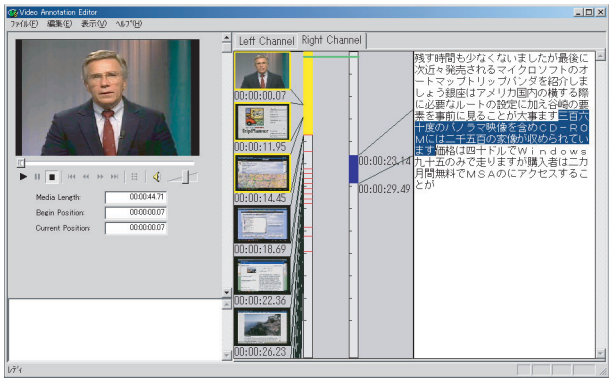
映像オブジェクトへのアノテーションの付与は、図 12 に示されるインタフェースを用いて行われる。このインタフェースでは、ビデオフレーム内の矩形で選択された領域のトラッキングやその領域で指定された映像オブジェクトへのテキストアノテーションや意味属性の付与を行うことができる。

9.3 アノテーションに基づく検索処理

生成されるアノテーションデータには、代表シーン画像やオブジェクト画像も含まれるが、これらは検索後の



左チャンネル



右チャンネル

図 10: 多言語ビデオトランスクリプタ

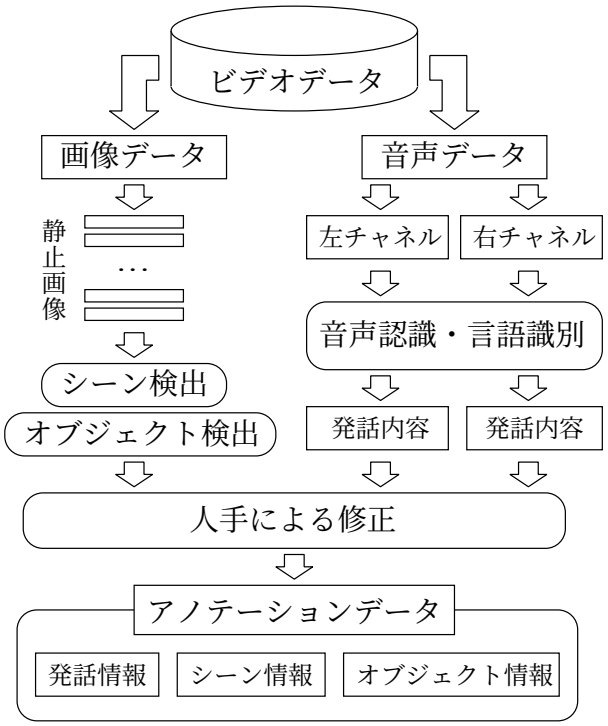


図 11: ビデオアノテーションデータ生成までの手順

図 12: 映像オブジェクトへのアノテーション

コンテンツ変換や要約時に使われ、検索時には利用されない。検索の原理としては通常のテキスト検索と同じであり、代表的なベクトル空間モデルを用いて類似度を計算する。

ビデオデータの場合は、類似度の高いデータの中から目的のシーン (タイムコード) を検出することが要求されるため、類似度計算はビデオファイル全体とビデオファイル中に含まれる全シーンに対して行われる。

ビデオファイル V_i が m 個のシーンから構成されるとする。

$$V_i = \{S_1, S_2, ..., S_m\}$$

このとき、各シーンの特徴ベクトルを $\overrightarrow{V_i(S_j)}$ とすると、ビデオ V_i の特徴ベクトルは、シーンの時間長とターム数によって正規化された次の式で表される。

$$\overrightarrow{V_i} = \sum_{j=1}^{m} \frac{l_j}{L} \cdot \frac{n_j}{N} \cdot \overrightarrow{V_i(S_j)}$$

(l_j: シーン時間長、L : ビデオ時間長、
n_j : シーン中ターム数、 N : 全ターム数)

各シーンの特徴ベクトル $\overrightarrow{V_i(S_j)}$ は、発話情報 $S_j(t)$ 、シーン情報 $S_j(s)$ 、オブジェクト情報 $S_j(o)$ 中に出現するタームのタームベクトルと、各々の情報に対する重み係数 α、β、γ によって以下の式で表される。

$$\overrightarrow{V_i(S_j)} = \alpha \cdot \overrightarrow{S_j(t)} + \beta \cdot \overrightarrow{S_j(s)} + \gamma \cdot \overrightarrow{S_j(o)}$$

$$(0 \leq \alpha, \beta, \gamma \leq 1)$$

タームベクトルを構成する各タームの重みは、$TF \cdot IDF$ 法 によって求められる。

検索は、まず全ビデオファイルに対してビデオの特徴ベクトルを用いて類似度計算を行い、次にその中の上位 R 件に含まれる全シーンに対して同様に類似度計算を行いランク付けする。最終的に、ランキング上位の r シーンを検索結果として出力する。

9.4 システムの問題点

多言語ビデオトランスクリプタにより、人間の介在する半自動的なビデオアノテーション作成の支援が可能になったが、アノテーションを付与する際の問題として、次の 2 点が挙げられる。

- 内容記述量、精度
- 内容記述コスト

これらは、処理の自動化と検索・要約等の実用性の意味において互いにトレードオフの関係にある。そこで、以下ではそれぞれの問題を解決する上で今後取り組む必要のある技術的課題について考察する。

9.4.1 内容記述量・精度

すでに述べたとおり、マルチメディアコンテンツは内容に基づく処理が極めて困難であることから、アノテーションツールを用いた内容記述は、ビデオ等のコンテンツの検索・要約を実現するために重要な役割を果たすと考える。しかし、内容記述が乏しかったり、機械処理に頼ることで内容記述の精度が悪いと、その後の検索や要約結果の精度を低下させる原因となる。

図 13 に示すグラフは、音声認識誤りをシミュレーションして、実際の検索精度にどの程度影響を及ぼすかを分析するとともに、検索において期待されるアノテーション情報の質を予測したものである。

具体的には、RWCP 検索・要約用ニュース音声データベース (192 記事、3,517 異なり単語)[16] を用いて、検索対象となる記事中に含まれる単語をランダムに除去した場合と、文書の特徴付けに有効でない単語を IDF(逆文書頻度) 値により優先的に除去した場合の 2 通りについて検索実験を行い比較した。

評価尺度には、再現率と適合率から求められる F-measure を用いた。入力クエリーにオリジナルの記事を用いており検索対象数も少ないことから、除去率が小さいときには両者に大きな差は見られないが、除去率 90%になるとその差は 40%程度になる。すなわち、検索に有意な 10%以上の単語を音声認識の脱落誤りによって失うと、検索精度に多大な影響があることがわかる。さらに、置換誤りや挿入誤りが起こると検索精度に対する信頼性は著しく低下する。

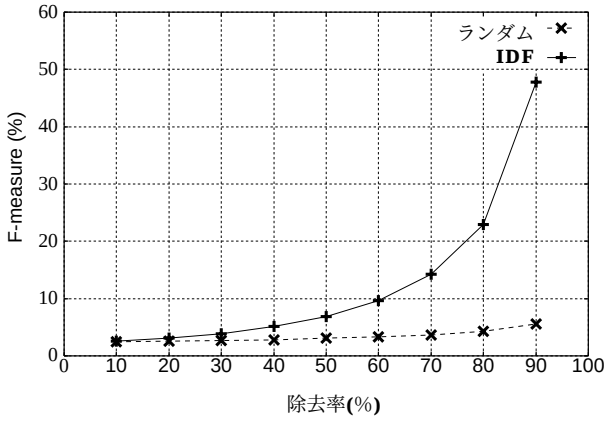


図 13: 単語除去率と検索精度の関係

そこで、内容記述量を向上させると同時に内容記述精度そのものを向上させる技術が必要である。前者については、アノテーションツールによる内容記述処理の自動化によってある程度は達成されているが、オブジェクト認識・トラッキング等の画像処理や発話内容に基づく談話構造化など、改良の余地はまだ多く残されている。後者については、発話文認識精度向上のための言語モデルの修正や、認識結果に対する事後処理としての認識誤り訂正処理が挙げられる。

特に、入力音声については、話題に応じた言語制約を加えることにより精度改善が期待される。音声対話システムの場合はリアルタイム性を重視するために、通常音声認識処理は 1 度しか行われないが、本研究のようなオフライン処理の場合は、音声認識と言語モデル修正を繰り返し行うことにより、認識結果の精度改善を図ることが可能であると考える (図 14)。言語モデルの修正には、認識結果以外に、クローズド・キャプション、Web テキスト、アノテーション情報等の外部知識が利用できる。

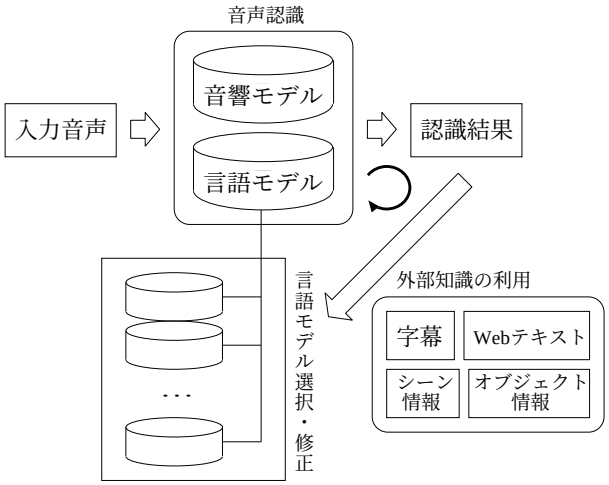


図 14: 話題に応じた言語モデルの選択・修正処理例

また、内容記述量・精度を客観的に示す評価尺度が必要である。コンテンツ時間長に対する内容記述量や、発話文認識結果の認識精度、構文としての正しさ、等を統一的な評価尺度で測った上で検索時にその評価量を導入することにより、検索結果に対する信頼性を向上させることが可能であると考えられる。

9.4.2 内容記述コスト

アノテーション付与において内容記述にかかるコストは、本システムにおける最大の課題である。内容記述にともなう機械処理の精度を極限もしくはタスク達成に必要な精度まで向上させることが最も重要であることは言うまでもないが、現実的に実用レベルのアノテーションを付与するためには人間の介在が不可欠であるため、人間がスムーズに作業を行うための入力支援やインタフェースの改良も必要である。

現時点の多言語ビデオトランスクリプタを用いて、Windows PC (CPU: PentiumIII 800MHz、Memory: 512MB) 上でアノテーションを行った場合にかかるコストを表 3 に示す。アノテーション作業は、まずツールの使用方法を説明した後にテストファイルを用いて 10 分間操作に慣れてもらい、その後、新しいビデオファイル (88 秒) に対して作業を行ってもらった。これを計 3 人分計測した。シーンとオブジェクトについては、作業者によって結果が異なるため、作業時間をシーン数、オブジェクト数で平均してある。

表 3: 内容記述コスト

	作業者			
分類	A	B	C	平均
(a)	58			
(b)	88	70	74	77
(c)	23	30	17	23
(d)	528	394	425	449
シーン数	5	4	4	
オブジェクト数	2	3	2	
(b-d)	1014	764	755	844
計	1072	822	813	902

(認識結果の単語正解率 78.5%, 正解精度 73.9%)

(a) 機械処理 (音声認識、シーン検出) 時間 [*sec*]

(b) シーン統合・内容記述時間 [*sec*]

(c) オブジェクト切り出し・内容記述時間 [*sec*]

(d) 発話内容修正時間 [*sec*]

表 3 より、音声認識精度が 70〜80%の場合、アノテーション作業を行う人間にかかる時間的なコストは、機械処理にかかる時間の約 15 倍、アノテーションを付与する対象データ時間の 10 倍程度であることがわかる。誰もが手軽にアノテーションを付与できるようにするためには、今後さらなる改良が必要であると思われる。

10 まとめと今後の課題

より実用的なセマンティック・ウェブに向けて、筆者の考えと提言を述べた。また、筆者のグループで研究開発しているマルチメディアコンテンツのアノテーションとその応用システムの概要について説明し、アノテーションに基づいたビデオ検索を実現する手法を具体的なシステムを例に提案した。また、現状のシステムの問題点として、アノテーションにおける内容記述量・精度と内容記述コストを挙げ、これら解決すべき問題について考察した。今後は、システムの問題点を改善しながら個々の技術を検討し、大量のデータに対して提案した検索手法の有効性を評価していきたいと考えている。

また、昨年度の TREC-2001 より Video Retrieval Track が導入され [9]、ビデオデータにおける類似検索の研究促進が期待されているので、こちらの動向にも注目しながら研究を進めていく予定である。

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、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から、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。また、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も重要である。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ことができるようになるだろう。

参考文献

[1] Chieko Asakawa and Hironobu Takagi. Annotation-based transcoding for nonvisual Web access. In *Proceedings of the Fourth International ACM Conference on Assistive Technologies (ASSETS 2000)*, pp. 172–179, 2000.

[2] Stevan Harnad. The symbol grounding problem. *Physica D*, Vol. 42, pp. 335–346, 1990.

[3] Koiti Hasida. Global Document Annotation, 2002. http://i-content.org/GDA/.

[4] Steven C. Ihde, Paul P. Maglio, Joerg Meyer, and Robert Barrett. Intermediary-based transcoding framework. *IBM SYSTEMS JOURNAL*, Vol. 40, No. 1, pp. 179–192, 2001.

[5] MPEG. MPEG-7 Main Page.

[6] Katashi Nagao, Shigeki Ohira, and Mitsuhiro Yoneoka. Annotation-based multimedia summarization and translation. In *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING-02)*, 2002.

[7] Katashi Nagao, Yoshinari Shirai, and Kevin Squire. Semantic annotation and transcoding: Making Web content more accessible. *IEEE MultiMedia Special Issue on Web Engineering*, Vol. 8, No. 2, pp. 69–81, 2001.

[8] Shigeki Ohira, Mitsuhiro Yoneoka, and Katashi Nagao. A multilingual video transcriptor and annotation-based video transcoding. In *Proceedings of the Second International Workshop on Content-Based Multimedia Indexing (CBMI-01)*, 2001.

[9] Text REtrieval Conference (TREC). TREC-2002 Video Track, 2002. http://www-nlpir.nist.gov/projects/trecvid/.

[10] W3C. XML Linking Language (XLink) Version 1.0, 2001. http://www.w3.org/TR/xlink/.

[11] W3C. Naming and Addressing: URIs, URLs, ..., 2002. http://www.w3.org/Addressing/.

[12] W3C. Resource Description Framework (RDF) Model and Syntax Specification, 2002. http://www.w3.org/TR/REC-rdf-syntax/.

[13] W3C. Simple Object Access Protocol (SOAP) 1.1, 2002. http://www.w3.org/TR/SOAP/.

[14] W3C. The Semantic Web Community Portal, 2002. http://www.semanticweb.org/.

[15] W3C. Web-Ontology (WebOnt) Working Group, 2002. http://www.w3.org/2001/sw/WebOnt/.

[16] 伊藤克亘, 田中和世, 中沢正幸, 岡隆一. ニュース音声コーパスの構築. 日本音響学会講演論文集, pp. 171–172, 1999.

[17] 岡隆一, 西村拓一. パターン検索のアルゴリズム・マップ –crossmediator を支えるもの–. 人工知能学会研究会資料, SIG-J-A101, pp. 1–6, 2001.

[18] 森靖英, 岡隆一. WWW 上の文書・画像混在データのクロスメディア検索. 人工知能学会研究会資料, SIG-CII-2001-MAR-06, 2001.

[19] 森靖英, 高橋裕信, 岡隆一. 画像と単語の相互検索手法. 人工知能学会研究会資料, SIG-CII-2000-NOV-17, 2000.

[20] 西崎博光, 中川聖一. 音声入力によるニュース音声検索システム. 電子情報通信学会技術研究報告, SP99-108, pp. 91–96, 1999.

[21] 西田昌史, 緒方淳, 有木康雄. 話者と発話内容の同時検索に関する検討. 人工知能学会研究会資料, SIG-CII-2000-NOV-12, 2000.

[22] 大平茂輝, 長尾確, 白井克彦. アノテーションに基づくビデオ検索システムの提案. 情報処理学会研究報告, SLP-43-6, pp. 33–39, 2002.