セマンティック・トランスコーディング
–Semantic Webのために今やるべきこと– *

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

次世代の高度な知識共有のインフラである Semantic Web 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、Semantic Web と同等の機能を有する仕組みを研究してきた。その経験に基づいて、Semantic Web 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う。

2 Semantic Webのためにやるべきこと

Semantic Web[7] 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」Semantic Web の問題点は以下の通りである。

1. RDF や OWL 等の Semantic Web 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。

*Semantic Transcoding: What we should do now for building the Semantic Web by Katashi Nagao (Center for Information Media Studies, Nagoya University)

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。

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。

3 アノテーション：デジタルコンテンツの階層化

図 1 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。

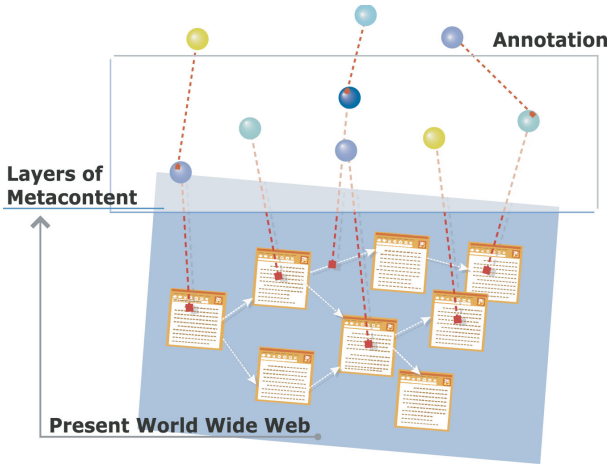

図 1: アノテーションによる Web コンテンツの拡張

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。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[4] 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[5] を用いることができるが、その方法論は

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 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。

4 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、それぞ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[3]。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。

5　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2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。

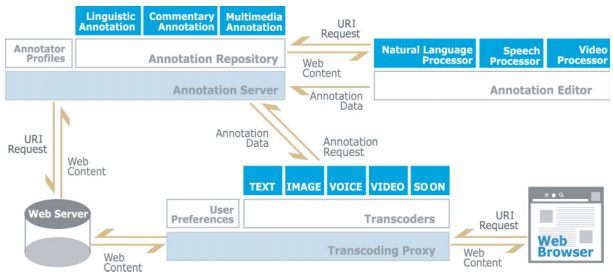

図2: セマンティック・トランスコーディングシステムの構成

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[2]。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なえる。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なえる。

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なう。

6 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」Semantic Web 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[8] 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、Semantic Web 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)[6] 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。

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] 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.

[3] 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.

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

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

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

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

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