

セマンティック・ウェブ技術を用いた Web コンテンツ配信*

佐伯嘉康[†]東京大学大学院新領域創成科学
研究科基盤情報学専攻[†]相田仁[‡]東京大学大学院新領域創成科学
研究科基盤情報学専攻[‡]

1 はじめに

Web 上には静的、動的に生成された Web サイト、Web ページなどの Web コンテンツが数多く存在しており、Web の利用者が Web コンテンツを発見または利用する場合、検索サイト、そして Web コンテンツのハイパーリンク構造に頼る作業が大きい。しかし、検索サイトは必ずしも全ての Web コンテンツを把握しているわけではなく、また、ハイパーリンク構造もまた全ての関連 Web コンテンツを網羅しているとは言えない。目的の Web コンテンツを効率的に利用するには、よりまとめられた形で発見出来る事が望ましい。

そこで、本配信手法では Web コンテンツの作成者、利用者が、Web コンテンツの内容に関するセマンティック Web 技術であるメタデータ (機械がデータを理解するための、データに関する記述) を付与する事で、内容に応じた Web コンテンツの集配を可能とする。

2 セマンティック Web 技術

2.1 セマンティック Web

セマンティック Web とは、1999 年の Tim Berners-Lee の WWW 国際会議の講演を端とし、Web 上のデータの全てに、機械による処理を可能にするための情報としてメタデータ (Metadata) を付与して、データの交換のための普遍的な媒体を形成し、データが持つ情報によってデータの (自動) 機械処理を可能にし、データの再利用性を高めるための次世代 Web の枠組みである [1]。現在、セマンティック Web 技術を実現するために様々な技術が提案され、組み合わせられている。

セマンティック Web は、データの関係と分類を改善するために必要なプロトコルと技術を確立する事により、システム間やデータ間関係を生成したり推論を行ったりする能力を強化する。

2.2 メタデータの記述

機械が Web コンテンツを理解するためには、Web コンテンツがメタデータを持つ必要がある。メタデータはデータに直接埋め込むか、データにリンクする形で存在し、データの情報について記述したデータである。セマンティック Web では RDF (Resource Description Framework) という W3C によって勧告されているデータモデルを使用して、メタデータを記述する。

RDF は、主語 (Subject)、述語 (Predicate)、目的語 (Object) の 3 つの要素のセット (三つ組, Triple) によって構成、記述されるデータモデルである。述語は RDF の中でタグ名として使われ、語彙ともいう。

RDF は現在、ソフトウェアの設定ファイルや RSS (RDF Site Summary) などで使用されている。しかし、Web 上の殆どのファイルが (X)HTML によって記述されたファイルであり、セマンティック Web 技術の普及のために全て RDF に変換する事は敷居が高く、Web コンテンツの利用環境の促進として現実的ではない。そこで、XHTML (Extensible HyperText Markup Language) と RDF の相互変換を可能にする手法として、XHTML にメタデータを埋め込む方法として RDFa が提案された。

2.3 メタデータ記述方式としての RDFa

RDFa (RDA/A, RDF with attributes) は XHTML 2.0 の構文に対応した文書内にアトリビュートとして、RDF と互換のあるメタデータを埋め込むためのデータモデルである。W3C (World Wide Web Consortium) が提案し、標準化に向けた議論が進められている。

XHTML のアトリビュートは、開始タグの中に図 1 の様に記述する。attribute が属性、name が属性値となる。

```
<TAG attribute="name">TEXT</TAG>
```

図 1 Attribute notation

* Web contents distribution based on Semantic Web technology

[†] SAEKI Yoshiyasu, Dept. Frontier Informatics, Graduate School of Frontier Sciences, The University of Tokyo[‡] AIDA Hitoshi, Dept. Frontier Informatics, Graduate School of Frontier Sciences, The University of Tokyo

機械が RDFa によって記述されたメタデータを処理する際には、属性と属性値を見て、データが何を示しているのか、何に関するデータなのかを理解する。属性値は名前空間 (Namespace) 内で定義された語彙 (述語)

集で定められている。従って、メタデータとして独自の語彙(属性値)を RDFa データ内で使用する場合は名前空間を用意しなければならない。名前空間は、名前空間識別子によって区分され、名前空間識別子は参照方法と参照場所を URI(Uniform Resource Identifier) の書式に従い記述する。

3 セマンティック・ウェブ技術を用いた Web コンテンツの配信

3.1 Web コンテンツに対するメタデータの付与

Web コンテンツの利用効率を上げるためには Web コンテンツの内容に即したメタデータの存在が必要不可欠である。そして、メタデータを付与するものとして、Web コンテンツの作成者及び積極的利用者は、内容に関するメタデータを付与するための、Web コンテンツの中身に対する理解を持っているものといえる。

Web コンテンツの内容のメタデータの語彙集として、ドキュメントの見出し/概要の公開のための語彙集である RSS と書誌情報の公開のための語彙集である Dublin Core を参考にし語彙集を準備する。

現在の検討において、少なくとも以下のメタデータ語彙を持つものとする。

- title(Web コンテンツのタイトル)
- creator/publisher/contributor(作成者/配布者/関係者)
- category(カテゴリ (folksonomy タグでは無く、限定された語から選択))
- description(要約, 抜粋)
- date(created/modified)(作成日時/更新日時)
- related(related contents, 関連/言及/参考 Web コンテンツ)

メタデータを付与した Web コンテンツの例として、「<http://example.org/webcontentsmetadata>」という URI によって示される名前空間で定義されているメタデータを付与した、「<http://example.com/rdfabook>」という URI によって示される本のレビューを記述した Web コンテンツの一部を図 2 に示す。

Web コンテンツに付与するメタデータは複数人の利用者により記述、共有される。現在の RSS やソーシャルブックマークサービスに見られる任意の文字列による分類、時系列に縛られた Web コンテンツの配信から、集合知によりながら揺らぎを抑えた形で内容に即した Web コンテンツの分類、配信を可能にする。

3.2 Web コンテンツの集約配信

Web コンテンツの利用は、検索エンジンの検索語の選出などにも見られる様に、ノウハウを必要とする面がある。そして、Web コンテンツの内容の理解に関し

```
<div xmlns:ex="http://example.org/webcontentsmetadata"
  about="http://example.com/rdfabook">
  <span property="ex:title">RDFa Book review</span>
  <span property="ex:creator">Alice</span>
  <span property="ex:publisher">Example Co.</span>
  <span property="ex:contributor">Bill</span>
  <span property="description">
    This is a "RDFa Book" Review by Alice.</span>
  <span property="ex:date">2007-02-02</span>
  <a href="http://example.com/xmlbook" rel="ex:related">
    http://example.com/xmlbook</a>
</div>
```

図 2 Web contents with metadata example

て利用者毎に差異が存在する [2]。また、論文の様にまとめられた形で存在する Web コンテンツは少なく、多くはメモや個人のブログの様な形で散在している。そこで、Web コンテンツをまとめる事によって、Web コンテンツをより体系的に効率的に利用する事を可能にする。

従来のハイパーリンクで構成された Web 構造の上では、Web コンテンツを利用するにはハイパーリンクを辿る方法でしか関連した Web コンテンツを参照/利用する事が出来ない。また、Web コンテンツの作成者は必ずしも全ての Web コンテンツにハイパーリンクを張っているとは限らない。

前節でメタデータの語彙として定めた“category”と“related”は、分類、参考コンテンツなどとして他の Web コンテンツとの関連を示している。従って、この関連によって繋がった Web コンテンツを利用の際に同時に配信する事は、Web コンテンツの利用の効率を高めると考えられる。

本研究において、Web コンテンツは 1 つのコンテンツサーバによって管理されており、一意の URI によって参照出来る。そこで、各 Web コンテンツが持つ関連を示すメタデータを利用し、関連付けられた複数の Web コンテンツを同時に見せる事で、利用効率を上げる。

4 おわりに

本稿では、Web コンテンツ配信に利用するためのセマンティック Web 技術として、メタデータ記述方式の 1 つ RDFa について解説し、それを配信に活用する方法について述べた。今後の課題として、配信技術を実現するシステムを開発し、有用性を検証する。

参考文献

- [1] 斎藤信男・荻野達也(監修), 財団法人 情報処理相互運用技術協会(編), “セマンティック Web 入門”, オーム社, 2004
- [2] 河合一夫, “協調活動における状況共有を表現するモデル”, 情報処理学会研究報告, 2006-DD-57, pp.23-28, September 2006