ISSN 2186-5647

6-25

セマンティック Web のための RDF/XML 文書作成支援システムの開発

日大生産工（院）　○高邑　誠
日大生産工　山下　安雄

1．はじめに

Web コンテンツを利用する際に、大半のユーザーは検索エンジンを使用する。しかし検索エンジンは、ユーザーが求める情報を素早く、正確に提供することは難しい。そこでコンピュータが効率よく情報を収集、解釈するために、Web コンテンツの内容に、意味を表す情報（メタデータ）を付加させて、機械に理解させやすくすることを目的としたプロジェクトのことをセマンティック Web という。セマンティック Web では、Web ページは XML によって記述した文書にタグを付ける。RDF とはメタデータを記述する枠組みであり、XML によって RDF を表現するための構文を RDF/XML と呼ぶ。RDF/XML 文書は Fig1 のように表す。又 RDF/XML 文書を主語、述語、目的語と 3 つの要素で表現したものを RDF トリプルといい、Fig.2 に例を示す。

```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description
rdf:about="http://www.nitidaitarou.com/">
<dc:title>日大太郎のブログ
</dc:title>
 <dc:creator> 日大　太郎
 </dc:creator>
 <dc:date>2011年
</dc:date>
<dc:coverage>日本
</dc:coverage>
</rdf:Description>
</rdf:RDF>
```

Fig. 1　RDF/XML Document

Fig. 2　RDF triple and graph

2．研究目的

現在、メタデータを Web コンテンツに記述する時には、人が RDF/XML 文書にメタデータを記述している。人によって記述されるメタデータ間では、メタデータを書く人によって、判断基準が違うため、一貫性に欠けやすく、メタデータの選択が不安定になりやすい。又、メタデータを記述する作業には、膨大な時間がかかる。そこでこれらの問題を解決するために、作業を機械に任せることで、メタデータの一貫性を保ちつつ、メタデータの記述における作業効率を向上させることを目的とする。

3．ダブリン・コア

ダブリン・コアはメタデータを記述する語彙の通称であり、WWW 上におけるリソースに関する情報を記述し、情報の探索や発見に役立てる目的で作られ、ネットワーク情報資源を記述するのに広く用いられていて、2003 年には ISO15386 として国際規格に採用され

RDF/XML Document Creation Support System for Semantic Web
Makoto TAKAMURA and Yasuo YAMASHITA

た。ダブリン・コアには、１５種類の基本要素と、基本要素より細かく分類された５５種類の精密化要素が定義されている。

４．実験手法

４.１　Webコンテンツから単語の抽出

任意のWebコンテンツの中から文章を選択し、単語の抽出を行う。

４.２メタデータの生成

抜き出した単語を分析し、プログラムに評価させて、ダブリン・コアの基本要素からメタデータを選択する。評価方法については後述する。

４.３RDF/XML文書の作成・検証

人が作ったものと、機械がメタデータを選択したRDF/XML文書を比較し、正誤率、作業効率、偏りなどを調べ、検証を行う。

４.４評価方法について

本研究では、点数評価法を用いる。オブジェクトを抽出し、対象とするオブジェクトの前後の単語の関係や単語の出現頻度などの情報などをプログラムに読み込ませ、条件が合うほど高い点数をつける。ダブリン・コアの基本要素に点数をつけていき、最終的に最も高い点数の基本要素を選択する。この方法は、点数の大きさによってオブジェクトとプロパティの距離を視覚化することができる。また間違えて記述した場合に訂正する場合にも、第二候補のプロパティが見つけやすくなり、修正にあたる時間が短くなるといったメリットがあると考える。

４.５ tf-idf法

tf-idfは文書の中に出現した単語がどの程度特徴的であるかを識別する指標であり、tf-idfによる重み付けを利用したtf-idf法の式を次に示す。

$$tf=f/\max(f) \qquad (1)$$

tfは単語の出現頻度で、文書内である単語がどれだけ使用されているのかを示す指標であり、同じ文書内に多く書かれている単語ほど値が大きくなる。fは単語の個数、max(f)は文書で出現する総単語数を表す。

$$idf=\log_2(n/df) \qquad (2)$$

idf は逆文書出現頻度で、ある単語がどれだけ別の文書で使用されているかを示す指標であり、他の文書内で書かれているほど値が小さくなる。n は総文書数、df は単語が含まれる文書数である。

$$W=tf*idf \qquad (3)$$

Wは重みであり、(1)の式と(2)の式を乗法して求める。このWの値が高いほどその単語が文書内のキーワードになる可能性が高くなる。

５．おわりに

今回は、検索エンジンの精度の向上のために、RDF作成支援方法について記述した。本研究は、作業時間の短縮やメタデータの一貫性を保つことや、オブジェクト間の距離の視覚化することが期待できる。しかし、機械が付けるメタデータの正誤率が低いとこれらの価値がなくなるため、適切な評価や、正確なメタデータの選択が必要となると考える。また選択したWebコンテンツによっても偏りがあると思われるため、サンプルが大量に必要であると考える。

参考文献

1) 今井憲一,　Webコンテンツからの RDF/XML文書の再構成方法と検証,平成20年度修士論文, (2009)pp.1~32.

2) 河本　穣, 単語の出現頻度を用いたドキュメントデータベースのメタデータの自動生成方式,電子情報通信学会第13回データ工学ワークショップ（DEWS2002）論文集, (2002)pp.2~10.

3)佐野公彦, 武井恵雄, 荒井正之, Web検索支援システムのためのRDFエディタの開発, 情報処理学会研究報告, (2003)pp.69~75.