

文書の意味的構造可視化のための文書タグ表現

Semantic tag description for document structure visualization

村山 正司 中村 裕一 大田 友一
Masashi Murayama Yuichi Nakamura Yuichi Ohta

筑波大学 機能工学系

Institute of Engineering Mechanics and Systems, University of Tsukuba

はじめに

大量の文書データから必要な情報を短時間で把握することは、未だに難しい問題である。本研究 [1] では文書内の意味的構造を形式的記述として付加したタグ付き文書から概念図を自動生成する手法 (図 1) を開発し、文書

メディアの概観性を高める枠組みを提案してきた。これまでは広範な内容記述の難しさから、限定された内容の文書にのみ適用してきたが、本稿ではより一般的な文書を記述するためのタグセットを定義し、それを新聞記事等に適用した。

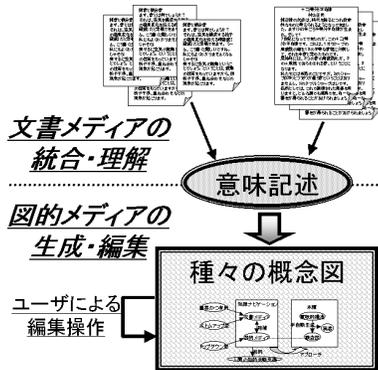


図 1: 文書の可視化による概観

文書内容を記述するためのタグセット

図的メディアへの変換を目的とした文書の意味的構造の記述は、XML1.0 に準拠した形式で行った。構造の記述については、構造を「型」と「性質」の2つの属性に分類して表現する。例えば次のようになる。

```
<a name="diet">国会は</a><a name="shu">衆議院</a>
と<a name="san">参議院</a>よりなる。
<slink type="order" axis="set" 1="diet" 2="shu san" />
```

1~2行目が要素の記述であり、3行目が意味的構造の記述となる。これにより”diet”と”shu”, ”san”と便宜上名付けられた「国会」と「衆議院」「参議院」が”order”という型の意味構造に対して”set”という性質の関係を持つことを表わす。構造の型を示す属性 type は、代数的な分類による次のような値をとる。

order	上下(順序)関係。反射律, 推移律を満たす。
eq	同位関係。反射律, 対称律, 推移律を満たす。
mod	修飾関係。対象に説明・属性を付加する。
binomial	上3者にあてはまらない2項関係。
complex	3項以上の特殊な関係。比較, 条件など。

構造の意味的性質を示す属性 axis は、次のような値をとる。

term(語義), time(時間), set(集合), attribute(特性), cause(因果), space(空間), inout(入出力), position(立場), content(内容), other(その他) などである。

実際の文書への適用

本稿の枠組みの上で、実際の新聞記事に対して以下のようにタグを付加した。リストの上半分がテキスト中の要素を宣言する部分であり、下半分がそれらの間の3つの関係を記述している。記述されている関係はいずれも修飾関係であり、上から順に「その他」、「内容」そして「特性」という意味的性質を表わしている。

```
<a name="attr1a">生保最大手</a>の<a name="nissei">
日本生命</a>保険と、<a name="attr2a">損保最大手</a>の
<a name="kaijou">三井・住友海上火災</a>は、<a name=
"kouhan">広範な分野</a>で<a name="teikei">提携</a>
する方針を固めた。
...(中略)
<slink type="mod" axis="other" 1="nissei"
2="kaijou" 3="teikei" />
<slink type="mod" axis="content" 1="teikei"
2="kouhan" />
<slink type="mod" axis="attribute" 1="nissei"
2="attr1a" />
...(後略)
```

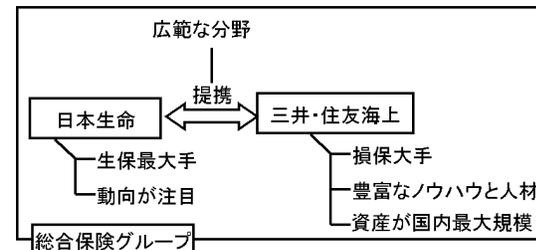


図 2: 目標とする概念図

このようにして記述された意味的構造から、本研究のこれまでの研究を利用して図 2 を生成するシステムを現在実装中である。

おわりに

一般的な文書に適用できる意味的構造のタグセットを定義し、これを実際の文書に付加してその大意を記述できることを示した。今後、手動に頼っているタグ付加を半自動化し、また意味的構造を可視化するシステムを完成させる予定である。

参考文献

- [1] 村山正司, 中村裕一, 大田友一: 概念図の自動生成による文書内容の可視化 - タグ付き文書からの自動変換 -, 信学会第 6 回知能情報メディアシンポジウム (1999).