概念図の自動生成によるタグ付文書の可視化

村山 正司† 中村 裕一†‡ 大田 友一†

†筑波大学 機能工学系
〒305-8573 つくば市天王台1-1-1
E-Mail: murayama@image.esys.tsukuba.ac.jp
‡科学技術振興事業団, さきがけ研究２１

あらまし: 現代社会では入手できる知識が膨大な量となり、それを人間にとってわかりやすく提示することが重要になっている。本研究では、最も基本的なメディアである文章と、直観的な理解プロセスを持つ図的メディアを相補的に用いた複合メディアによって知識を有効に提示することで、人間に対する知的活動支援を目的とする。そのために文章のもつ意味的構造に着目し、それが明示的に埋め込まれたタグ付き文書から概念図を生成することで複合メディアを提示する枠組みを提案する。本稿では、タグ付き文書に記述された意味的構造と概念図との対応について検討し、それを用いた概念図の自動生成について述べる。また計算機上での実験によってその妥当性を検証した。

キーワード: 概念図生成, ハイパーメディア, 知識表現, タグ付文書

Automatic Diagram Generation from Tagged Text

Masashi Murayama† Yuichi Nakamura †‡ Yuichi Ohta†

† Institute of Engineering Mechanics and Systems, University of Tsukuba
1-1-1 Tennoudai,Tsukuba, 305-8573,Japan
E-mail: murayama@image.esys.tsukuba.ac.jp
‡ PRESTO, Japan Science and Technology Corporation(JST)

Abstract: We often need diagrams for explanation, though a text is the most powerful and effective tools for communication. For this purpose, we propose a novel scheme for diagram generation, in which the semantic structure of a tagged text is effectively translated and linked to the text.
First, we analyzed the semantic correspondence between diagramatic expression and semantic the GDA tag sets. Then, we propose our framework for automatic translation from tagged text to diagrams.

key words: diagram generation, hypermedia, knowledge representation, tagged text

1　はじめに

近年、Internet や公共利用できるデータベースの発達に伴い、大量の情報を誰もが容易に入手できるようになってきた。そのような状況の中では情報過多となりやすく、本当に必要な情報を短時間で把握することが難しくなってしまう。

そのため情報の収集・蓄積・流通・提示を計算機によって効率的に自動化し、理解支援・伝達支援を行うことが必要である。特に、情報や知識を人間にとってわかりやすく表現することが重要である。しかし、単一メディアのみでは伝達効率が悪く、複数のメディアを相互補完的に利用した複合メディアを用いることで、単一メディアのもつ問題点を克服することが必要となる。

そこで我々は文章の内部構造から概念図を自動生成する手法を提案してきた [1][2]。文章と概念図を併用し、相互補完することによって人間にとって理解しやすい複合メディアとすることを目的とする。また、文章と概念図双方のメディアをハイパーリンクにより有機的に統合することにより、概要把握と詳細理解の双方を容易にしたハイパーメディアの実現を目指している。

本稿では、これまで十分に議論してこなかった文章の意味構造を入力する有効な方法について述べる。具体的には、意味的構造が記述されているタグ付き文章を利用することによりその問題を解決する方法を検討した。また、計算機に実装して概念図生成実験を行い、その有効性を確認した。

2　文章の意味的構造の可視化

本研究は文章メディアの内容の説明・補足を図的メディアによって行うことを目的としている。そのような図的メディアを生成するためには、図 1 のように大きく分けて次の二つの問題を解決する必要がある。

(a) 文章から意味的構造を抽出・解析する自然言語処理、又はユーザから直接入力するためのインタフェースの問題

(b) 意味的構造から図的メディアを生成し、ユーザの満足する図を得る問題

これらの問題に対して、本研究では以下のアプローチをとった。

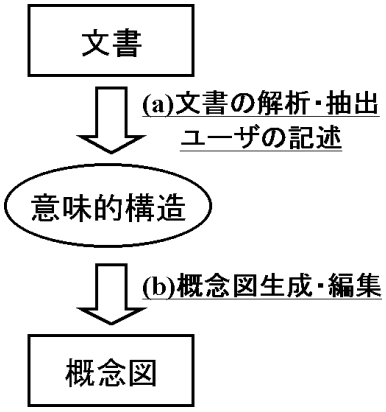


図 1: 文章可視化の概要

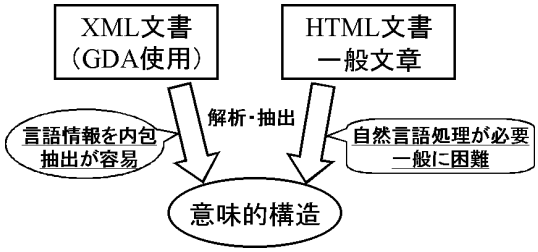


図 2: 意味的構造の解析・抽出

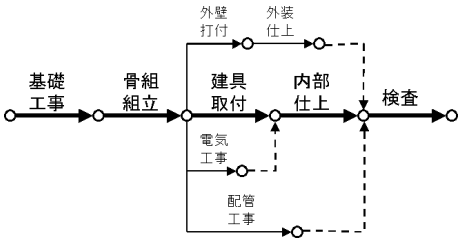
文章の意味的構造の抽出

文章から機械可読な形式の意味的構造を入力する必要があるが、これは簡単な問題ではない。そこで本研究ではこの問題を直接扱うのではなく、まずタグ付き文書について、その実現可能性を検証した。これについては４章で述べる。

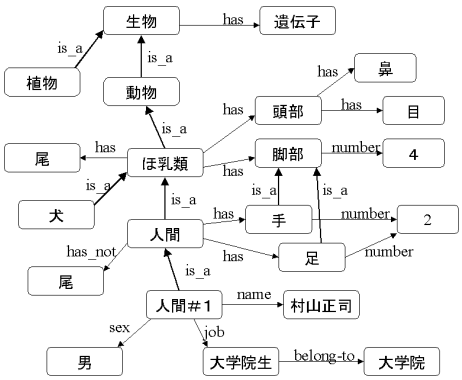
本研究で用いる GDA[3] は言語学的な構造や属性を記述できるように定義されており、概念図を生成するために必要な情報の多くを記述することができる。また、汎用的なデータ記述言語の XML の拡張であることから、それらの構造情報は容易に計算機で認識できる。このように GDA 等を用いたタグ付き文書を蓄積する研究は既に始まっており、今後はこのような形で文書が蓄積されていくことが期待されている。

概念図の生成

図的表現では、空間的な構造によって何らかの知識を表現する。しかし、その表現方法には様々なものがあり、その質の良し悪しによって概略把握の容易さ、理解効果の向上の度合いが決まる。そのため、我々は典型的な空間構造により表現できる意味



(a) 良い図の例（概略が掴める）



(b) 悪い図の例（焦点が判らない）

図 3: 図の例

的構造と文章の意味的構造との間の対応関係を調査し、その典型的な利用方法を提案している。これについては３章で述べる。

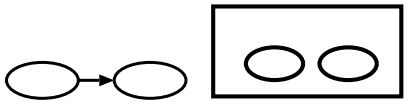
また、文章の意味的構造から概念間の関係構造を生成し、そこから概念図の空間構造を生成する多段階の変換手続きを設定した [1]。これについては５章で説明する。

さらに、自動生成される図では人間の主観にそぐわない場合がある。そこで人手による編集操作の支援を可能にした。そうした人間とシステムのインタラクションを通じて、より良い概念図の生成を行う。これについては６章で実際に例証する。

3　概念図の意味構造と構成規則

概念図の形式には様々なものがある [4]。例えば図 3(a) は手順を表わすアローダイアグラムであり、図 3(b) は多数の概念間の関係を表わす意味ネットワークである。これらの図には、人間にとってのわかりやすさにおいて大きな違いがある。

図 3(a) は重要な手順が強調され、クリティカルパスを把握することができる、良い図であると言える。それに対して、図 3(b) は、計算機内部の知識表現としては有用であるが、人間に提示するための表現としては繁雑であり、一目で内容を理解することは難しい。このように、判りやすさを示す記号的機能は伝達メディアとしての概念図にとって重要である。そのため、生成する概念図の形式を適切に選択することが重要な問題となる。



(a) 連結図　　(b) 領域図

図 4: 概念図

表 1: 文章の持つ意味的構造の大分類

順序関係	順序、系列、軸などの関係
包含関係	階層や上下関係を表わす関係
同値関係	同値・同等、並列を表わす関係
修飾関係	説明や属性を付加する関係

本研究ではまず、文章に近い意味的構造を有している２種類の基本的な概念図を扱うこととした [1]。構成要素同士を矢線要素で連結した連結図と、閉曲線により他の構成要素を囲む領域図である。これらを図 4 に示す。

この二つの概念図がもつ意味的構造を次に挙げる。

連結図 連結による同値・並列構造、有向矢線による順序構造

領域図 閉領域による包含構造、階層構造

文章の持つ意味的構造のうち、これらの図形構造で表現可能な関係を表 1 に挙げる。また、自然言語の持つ多様な意味的構造を表現するため、表 2 に示すような属性情報をこれらの関係に付加する。表 1 で挙げた構造と概念図の空間構造は、次のように対応付けられる。

順序関係: 連結図の持つ構造と対応する。
図形同士を矢線で連結し、順序と軸を示す。

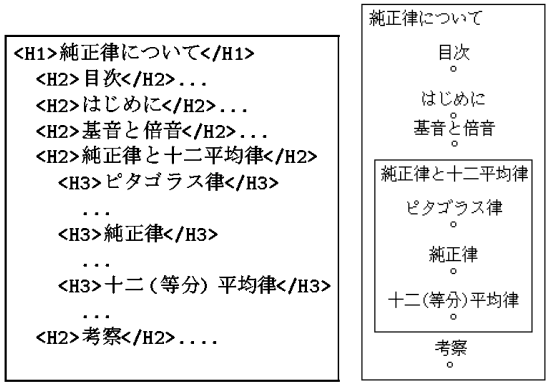
包含関係: 領域図の持つ包含構造と対応する。
上位概念が下位概念を囲むような空間構造。

同値関係: 順序関係に準じる。

修飾関係: 順序関係に準じる。

表 2: 主な属性

時間	時間軸、時系列を示す属性
因果	原因・理由を示す属性
空間	実空間での物理的な属性
入出力	過程や原材料と生成物の関係を示す属性
話題	話題の流れ上にある属性
集合	組織や集合論上の関係を示す属性



(a) HTML 文書　　(b) 概念図

図 5: 見出し構造の可視化

この対応関係を図の空間構造の構成規則として設定し、入力した文章の意味構造に適用することで、概念図の自動生成が可能となる。

4　タグ付き文書からの変換

現在、タグ付き文書としては HTML が最も広く使われているが、HTML を構成するタグセットは主としてハイパーリンクの定義や装飾・整形などの視覚効果を目的としたものである。それらの中には見出し要素（<H1>...</H1>）や箇条書きなど、限定された構造を定義するタグは存在するが、意味的な構造や複雑な論理構造を記述することはできない。例えば図 5(a) に示す HTML タグ付き文書には見出し間に階層構造が存在し、その構造を包含関係で表現した場合、図 5(b) に示すような概念図となる。しかし現状ではこのように単純な構造以外を HTML から取得することは自然言語処理を要するため、実用には困難が伴う。

GDA は、XML(eXtensible Markup Language) を拡張するタグセットの一種であり、自然言語の持つ意味をタグの形式で明示することが目的である。人間があらかじめ電子化テキストに言語学的なタグを付加しておくことで、それらテキストの計算機による処理を簡単かつ高精度にする。本研究では、GDA タグセットのうち概念図の生成に適した部分について着目した。

```
<su>
  <np id="clnt" gol="serv">
    クライアント
  </np>から
  <np id="serv" src="clnt" pat="req">
    サーバ
  </np>に
  <np id="req">要求</np>が送られる。
</su>
```

図 6: GDA による文内構造記述例

4.1　GDA による文章意味構造の記述

本研究で対象とする文章中の意味的構造には、大きく分けて文内構造、文脈・文書構造の 2 つがある。

文内構造は単語や語句間に存在する構造で、動詞の語彙的な意味構造や格構造、並列句などがある。例えば次のような構造がある。

- 概念の類似性や同一性
  例）A、B、C 等が挙げられる。
- 論理的・物理的な関係や構造
  例）A とは B の一種である。
  例）A、B、C と順に行う。
  例）A から B に C が送られる。

ここで、物理的な移動を表わす「クライアントからサーバに要求が送られる。」という文章を GDA タグセットによって表現した例を、図 6 に示す[1]。GDA においては、表 3 に示すタグによりタグ付けされた要素に対して、atts の部分に識別子の指定を記述することができる。また後述する関係属性を同様に記述することで、要素間の 2 項関係を表現することができる。

「<np id="...">...</np>」という記述は「クライアント」「サーバ」「要求」の 3 単語がそれぞれ名詞であることを示しており（表 3）、また各単語に識別子を設定している。「クライアント」を示すタグ内には、「サーバ」が何らかの変化の終末位置 (gol) であることが記述されている。また、「サー

[1] 見やすくするため、空白及び改行を施している。以下同じ。

表 3: GDA 構成要素タグ

<su atts>...</su>	文
<ss atts>...</ss>	複数の文
<np atts>...</np> <n atts>...</n>	名詞句
<namep atts>...</namep> <name atts>...</name>	固有名詞句
<vp atts>...</vp> <v atts>...</v>	動詞句
<ajp atts>...</ajp> <aj atts>...</aj>	形容詞句
<adp atts>...</adp> <ad atts>...</ad>	副詞・連体詞・接続詞・助詞
<seg atts>...</seg>	その他の構成素

バ」を示すタグ内には、開始位置 (src) が「クライアント」であること、被動作対象 (pat) が「要求」であることが記述されている。

文脈・文書構造は以下のように扱う [5]。例えば典型的な起承転結の構成を持つ次の文章の構造について考えてみよう。

1. 大阪本町糸屋の娘。
2. 姉は十六、妹は十五。
3. 諸国大名は弓矢で殺す。
4. 糸屋の娘は目で殺す。

ここで第 1 文と第 2 文の間には展開の関係が存在し、第 3 文と第 4 文には対比の関係が存在する。また、第 2 文と第 3 文の間には話題転換の関係があり、第 1 文および第 2 文と第 4 文の間に展開の関係がある。この 4 文を GDA を用いて明示的に記述したのが図 7 である。「<su id="...">...</su>」という記述は文に識別子を設定している。タグ中に存在する「ela="s2"」などの記述は、上述の文脈構造を表現している。このような構造を単純な図で説明すると図 8 のようになる。以上のようにして、GDA を用いて文内構造及び文脈構造が表現できる。

4.2 GDA の関係属性

以上のように、文脈構造、文内構造の両者をともに GDA タグの関係属性によって記述することができる。この関係属性とは、GDA 中のあるエレメントが他のエレメントとどのような関係を持っているかを示す属性値のことである。

例えば次のような GDA タグ付き文書を例にとる。

```
<su id="s1" ela="s2">大阪本町糸屋の娘.</su>
<su id="s2" otr="s3">姉は十六、妹は十五.</su>
<su id="s3" cntrst="s4">諸国大名は弓矢で殺す.</su>
<su id="s4" agt="s1 s2">糸屋の娘は目で殺す.</su>
```

図 7: GDA による文脈構造記述例

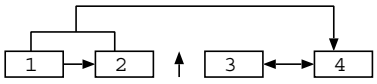


図 8: 「糸屋の娘」の文脈構造

```
<np sup="prima">ヒト</np>は
<np id="prima">霊長類</np>である。
```

「ヒト」という名詞句をマークアップしている一つ目のタグでは、関係属性 sup に属性値 prima が与えられており、「ヒト」の上位概念が識別子 prima で示される意味的要素であることを示している。そして、二つ目のタグでは「霊長類」という名詞句に prima という識別子を与えている。よって、このタグ付き文書からは、「ヒト」の上位概念は「霊長類」である、という意味的構造を抽出できる。

GDA で定義されている関係属性のうち、本研究に関連が深いものを、表 4 に挙げる。我々の枠組みで概念図を生成するためには、関係属性における分類と表 1 に示した 4 種類の意味構造とがおおむね対応している必要がある。それを示したのがこの表 4 右側である。この対応関係を上で示した例に適用すると、集合属性を持つ「ヒト $\subseteq$ 霊長類」という関係の記述を得る。

表 4 の関係より、GDA の関係属性と概念図の形式の間の対応関係が表 5 になる。

5 概念図の半自動生成手法

概念図生成の流れについて概説する。本研究では概念図の生成過程において複数のデータ構造を用意する。すなわち、入力データを多段階変換してゆくことにより、概念図を生成する。また、各データ構造間の変換は、各段階で変換ルール群を適用することで実現した。

表 4: 関係属性と意味記述との対応

GDA 関係属性		意味記述	
関係属性名	説明	関係名	属性
ini	時間的始点	順序	時間
fin	時間的終点		
ela	詳説、展開		話題
src	初期状態		空間
gol	終末状態		
mat	素材、材料		入出力
res	結果、産物		
cau	原因		因果
pur	目的		
cnd	事象の条件	修飾	
bas	根拠		
agt	行為者		話題
pat	被動作対象		
ben	行為受益者		
cnt	内容		
sum	要約		
eg	例示		
eq	等価、照応	同値	集合
sub	下位、部分	包含	
sup	上位、全体		

表 5: 関係属性と図の空間構造との対応

対応する図	文章の意味構造	GDA 関係属性
連結図	順序関係	src,gol,cau 等
	同値関係	eq
	修飾関係	cnd,bas,agt 等
領域図	包含関係	sup,sub

5.1 概念図生成のためのデータ構造

概念図の生成過程において扱うべき表現が複数存在する。まず、必要な表現は入力する知識の表現だが、GDA 文書から取り出した記述を、以下のような形式にする。

意味記述

意味的構造を、概念図の構造に合わせた形式として記述するために、宣言的記述を用いたデータ構造を用意した。本稿ではこれを意味記述と呼ぶ。具体的には図 9 に示すように、意味的要素および意味構造の宣言に接頭辞付パラメータリストをとって意味構造の記述を行う。

「要素」という接頭辞を持つ 1 行目から 3 行目は、A、B、C という識別子を持つ意味的要素を宣言している。4 行目が意味構造の記述であり、識別子 A で示される意味的要素が同じく B、C で示される要素の上位階層であるという宣言を行っている。

```
要素 (A,LABEL="XML")
要素 (B,LABEL="XML 宣言")
要素 (C,LABEL="DTD")
階層 (A,(B,C))
```

図 9: 意味記述の例

図フレーム

本研究では、概念図の図的構造・空間的構成を記述するデータ構造としてフレームを用いた。本稿では図フレームと呼ぶ。各フレームがそれぞれ一個の図形要素に対応する。ここでの図形要素を以後プリミティブと呼ぶ。

フレームの構成要素であるスロットにプリミティブの属性情報が保存される。位置や大きさ、色や意味的属性などである。他のプリミティブとの空間的関係情報をも同時に保存する。

関係ネットワーク

意味構造とは文章メディアの意味的表現であり、図フレームは図的メディアの内部表現であるため、これらの間に意味的変換が必要となる。本研究では複数の意味記述を統合するためのデータ構造として関係ネットワークを用いた。文書の意味記述における要素をノードとし、それら要素間の関係をリンクとして表現する意味ネットワークである。

このように意味記述から関係ネットワークを構成し、関係ネットワークから図フレームを構築するという過程を経ることで、概念図を無理なく生成できる。

5.2 データの変換

意味記述から関係ネットワークを構成する変換手続きは次のようになる。

1. 意味記述中の意味的要素に対応するノードを生成する。
2. 意味記述中の意味構造を入力し、対応するノード間リンクを生成する。

基本的にはこの二つの単純な手続きだが、ノード生成の際に意味記述での意味的属性を継承させたり、複数の記述を統合・圧縮してまとめる処理も行われる。関係ネットワークから図フレームを構築する変換手続きは下のようになる。

```
<su>
  <n id="serv">サーバ</n>には
  <n id="www-serv" sup="serv">
    WWW サーバ
  </n>、
  <n id="ftp-serv" sup="serv">
    FTP サーバ
  </n>がある。
</su>
```

図 10: 包含関係を示す GDA タグ付き文章

```
要素 (A,LABEL=”クライアント”)
要素 (B,LABEL=”サーバ”)
要素 (C,LABEL=”要求”)
要素 (D,LABEL=”WWW サーバ”)
要素 (E,LABEL=”FTP サーバ”)
推移 (A,B,C)
階層 (B,(D,E))
```

図 11: 複雑な意味記述

1. 関係ネットワーク上の各ノードから、図フレームの原型を構成する。
2. 関係ネットワークのリンクに保存されているノード間の関係情報から、図フレーム間の空間的位置関係を生成する。その際に適用される構成ルールは以下のようになる。

   順序関係: 概念要素同士を連結する連結要素は概念要素に接触していなければならない。一連の関係にある要素群は、初期状態では一直線上に並べられねばならない。

   包含関係: 包含要素は被包含要素群を囲む閉曲線でなくてはならない。

   同値関係: 順序関係に準ずる

   修飾関係: 順序関係に準ずる

これらの手続きを経た後、図フレームの座標値スロットに具体的な値を計算し書き込むことで概念図の生成が終了する。

6 実験例

6.1 図の生成実験

まず実験のために用意した意味記述からの概念図の生成例を示す。前述の図 6 は「クライアントからサーバに要求が送られる。」という順序関係を表現した GDA タグ付き文章だが、それに加えて図 10 のように階層関係を持つ GDA 文書を実験用のデータとする。これらをまとめて意味記述に変換すると、図 11 が得られる。これをシステムに入力すると、図 12 を得た。この図から、順序関係と包含関係が複合した構造を一目で見ることができる。

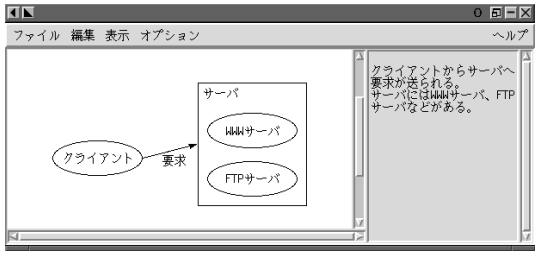


図 12: 生成された複合図

```
<su>
  <np id="hito">ヒト</np>
  の文化は... 二つの
  <n id="events">出来事</n>
  によって飛躍的な発展を成し遂げてきた.
</su>
<su>
  その一つは,
  <np sup="hito" id="homo">
    homo sapiens
  </np>
  が生まれるまでの
  <np id="prima" sub="hito">霊長類</np>
  の進化の過程で生じた,
  <np id="event1" sup="events">
    巨大な高次連合野の形成
  </np>
  である.
</su>
<su>
...（中略）... 第二の革命的事件は
  <np id="event2" sup="events">
   文字の発明
  </np>
  である.
</su>
...（後略）
```

図 13: GDA タグ付き文書

次に、実際の文章からタグ付けを行った GDA 文書を用意した。しかし、そこには概念図の利用が有効な関係属性は少なかったため、それらの中で代表的な語句間の論理的な階層構造のみを抜き出したのが図 13 に示す GDA タグ付き文書である。それを表 4 の対応関係に基づいて意味記述に変換した結果が図 14 である。その意味記述をシステムへ入力すると、概念図自動生成の結果として図 15 を得た。得られた図から、上記文章に含まれる論理的な階層構造を一目で見ることができる。

```
要素 (hito,LABEL="ヒト")
要素 (homo,LABEL="homo sapiens")
要素 (prima,LABEL="霊長類")
要素 (event,LABEL="出来事")
要素 (event1,LABEL="巨大な連合野の形成")
要素 (event2,LABEL="文字の発明")
階層 (hito,homo)
階層 (prima,homo)
階層 (prima,hito)
階層 (event,(event1,event2))
```

図 14: 意味記述

図 15: 実際の GDA 文書からの生成実験

6.2 図に対する編集操作実験

次に実験用の意味記述を用いて編集操作の実験を行った。まず複数の順序関係が複合した意味構造の記述を入力した際に、図 16(a) に示す図形が得られた。この図は人間にとって満足できるものではないので、人手による編集操作を加える。

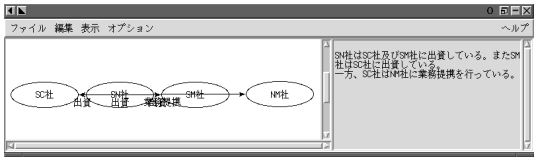
そこで図中の４つの楕円プリミティブをドラッグ&ドロップ操作により移動させた。その過程では「矢線プリミティブは接触すべきプリミティブと常に接触し続ける」というルールが適用される。その結果、動かした楕円プリミティブに関係する矢線プリミティブが編集ルールにより変位・変形され、図 16(b) が得られた。

このように、少ない手順で比較的良い図が得られる。編集操作で図の意味的な構造を損なわないようにシステムが人間の補助を行った結果であると言える。

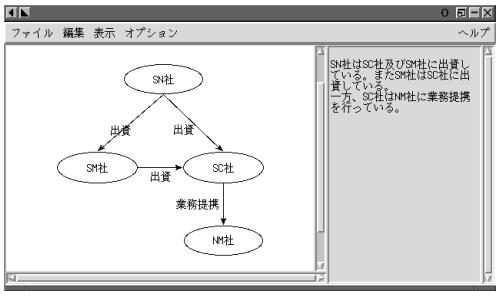
7 おわりに

タグ付き文章で表現された文章から概念図を生成する手法について提案した。GDA によってタグ付けされた文章の示す意味的構造と、概念図との間の対応関係を調査し設定した。概念図の生成過程で処理すべき構造に適したデータ構造を多段階変換する。そして実験の結果、概念図の生成が可能であることが実証された。

今後の課題としては、多段階変換の過程で推論などの知識処理を実装することで、より複雑な意味的構造を可視化することが挙げられる。



(a) 編集前の状態



(b) 編集後の状態

図 16: 編集実験

参考文献

[1] 村山正司, 中村裕一, 大田友一：知識ナビゲーションのための概念図の自動生成, OFS 99-21, 電子情報通信学会研究報告 (1999).

[2] 中村裕一, 村山正司, 大田友一：図的メディアと言語メディアの統合による知識の解析と提示, 知能情報メディアシンポジウム予稿集 (1998).

[3] 橋田浩一：GDA:意味的修飾に基づく多用途の知的コンテンツ, 人工知能学会誌, Vol. 13, No. 4, pp. 528–535 (1998).

[4] 出原栄一, 吉田武生, 渥美浩章：図の体系〜図的思考とその表現〜, 日科技連 (1986).

[5] 永野賢：文章論総説：文章論的思考, 朝倉書店 (1986).