人工知能学会研究会資料
SIG-SW&ONT-A301-05

オントロジーの自動構築に関する基礎的研究

内田英里　石野武志
関西大学総合情報学部古田研究室
〒569-1095 大阪府高槻市霊仙寺町 2-1-1
fa90029@edu.kutc.kansai-u.ac.jp

概要

現在，オントロジーは人間の手作業によって構築されることが一般的であり，この作業は膨大な時間と専門家の知識を要する．これらの問題点を克服するためには現在，存在するリソースを用いてオントロジーの自動構築を行う方法が考えられる．そこで本稿では土木分野に焦点をあて，辞書ファイルや Web テキストに対してルール解析を行うことによって，用語だけを体系化したオントロジーであるシソーラスの自動構築を行い，結果と考察を述べる．

１．はじめに

近年，インターネットは，情報の収集，蓄積，処理，利用に様々な技術進化をもたらし，企業，商業，マーケティング，財政，出版，教育，研究，開発などの様々な分野に新たな可能性を提供している．一方，知識システムの巨大化／複雑化が進み，利用者は，目的に応じた情報の収集，選択，統合化を行う必要がある．そこで，情報化社会において，情報を単なる記号として処理する技術にとどまらず，知識を利用した情報処理技術，特に知識共有／再利用を目的に，情報が表す内容を知識として扱う必要性が高まってきている．

この様な背景のもと，対象世界の知識を体系化するオントロジーに注目が寄せられている．オントロジーは，人間の手作業によって構築されることが一般的であるが，この作業は，膨大な時間と専門家の知識を要する．本研究では，土木分野に焦点をあて，辞書と WEB を用いて，オントロジーの自動構築を試みた．

２．オントロジーとは

オントロジーとは「知識システムを構築する際の構成要素として用いられる基本概念／語彙の体系」を示し，知識ベースを構築する際の背景となるバックボーン情報を提供するため，知識の共有／再利用に大きく貢献すると考えられている技術のことである．本研究では，オントロジー体系を，単語の同義語，狭義語，広義語，関連語を定義した体系である「シソーラス体系」として構築を試みる．

３．ルール解析

ルール解析は，日本語独自の文法を基に，広義語，狭義語，同義語の分析，予測を行う．

３．１　日本語の文の基本的な構成

日本語の文では，基本的に，文の中で主語の要素として働く『主部』と述語の要素として働く『述部』から構成される．言い換えると，『主部』とは，文の主題になる部分であり，『述部』は『主部』について述べる部分である．

本研究では，この主部と述部に注目し，主部からは，『主語となる名詞(以下 SS)』と『主語となる名詞の次にくる助動詞もしくは助詞(以下 SVi)』を，述部からは，『動詞，助動詞もしくは助詞の直前にある主語を保護する働きをする名詞(以下 PS)』と『その直後にある助動詞もしくは助詞(以下 PVi)』を抜き出し，SVi と PVi を基に，SS と PS の関係（広義語，狭義語，同義語）の分析をする．また，SS，SVi，PS，PVi の

組み合せは，いくつものパターンが考えられる．

３．２　A[狭義語]は　B[広義語]だ

『A[狭義語]は　B[広義語]だ』は，次のような構成のものを示す．

『A（名詞）＋は（助詞-係助詞）＋～＋B（名詞）
＋だ（助動詞)』
↓
『A』は『B』の狭義語である
『B』は『A』の広義語である

３．３　A[広義語]には　B[狭義語]がある

『A[広義語]には　B[狭義語]がある』は，次のような構成のものを示す．

『A（名詞）＋に（助詞-格助詞）
＋は（助詞-係助詞）＋～＋B（名詞）
＋が（助詞-格助詞）＋ある（動詞-自立)』
↓
『A』は『B』の広義語である
『B』は『A』の狭義語である

３．４　A[同義語]は　B[同義語]だ

『A[同義語]は　B[同義語]だ』は，次のような構成のものを示す．

『A（名詞）＋は（助詞-係助詞）＋～＋B（名詞）
＋だ（助動詞)』
↓
『A』は『B』の同義語である
『B』は『A』の同義語である

３．５　実装方法

ルール解析処理においては，本章で述べたルールにマッチした文から単語と単語の関係を抜き出す．ここでは，ルール解析の実装手段を簡単に述べる．

①一文ずつに分割
②一文を形態素解析ソフト『茶筅』にかけ，形態素の詳細情報を取得
③茶筅の結果の詳細情報を，CSV 形式に変換する
④変換したCSVを正規表現を用いてルール解析し，関係を取得する

４．関連語解析

本研究で行った関連語解析について述べる．まず，関連語関係を分析する上で，単語の出現頻度と単語の出現位置をもとに推考していく．

４．１　単語の出現頻度による評価方法

単語の出現頻度は，以下のようにしてもとめる．分析する文章の総テキスト数を N, 総テキストのうち単語 i が出現したテキスト数を $df(i)$, テキスト d に出現した単語 i の回数を $df(i)$, テキスト d に出現する総単語数を $c(d)$とする．

まず，一テキスト d あたりの単語数に関する重さ $dk(d)$を

$$dk(d)=Log(cd(d)+\alpha)$$

(式　4.1.1)

で求める．α を加算するのは $dk(d)$による変化率を下げるためである．

次に，テキスト d における，単語 i と単語 j に関する相関値を求める．単語 i が出現した際に単語 j が出現する度合いを

$$\min(df(d,i),df(d,j))/df(d,i)$$

(式　4.1.2)

とする．しかし，一テキストにおける単語数が多ければ，同時に出現する確率も高くなるので，これを $dk(d)$で割り，割った値をテキスト d における単語 i と単語 j の相関値 $fp(d,i,j)$とした．

$$fp(d,i,j)=(\min(df(d,i),df(d,j))/df(d,i))/dk(d)$$

(式　4.1.3)

そして $fp(d,i,j)$を全てのテキストに対して求め, 平均する．

$$fp(i,j)=\sum fp(d,i,j)/N$$

(式　4.1.4)

この値を用いて，以下の式で，頻度による相関値 $frel(d,i,j)$を求める．

$$frel(i,j)=fp(i,j)\times(df(i,j)/df(i))$$

(式　4.1.5)

４．２　単語の出現位置による評価方法

次に，出現位置を基とした単語と単語の相関値を求める．まず，単語 i と単語 j の距離を配列に代入し，$da(i,j)[]$とする．まず，それぞれの距離に対する評価値を，

$$1/Log(da[k]+\alpha)$$

（式　4.2.1）

とし（但し，βは係数），

$$\sum(1/Log(da[k]+\beta))$$

（式　4.2.2）

により，テキスト d における単語 i と単語 j の距離に対する評価値の合計を求め，最後に単語 i と単語 j の距離 $betweenCount(i,j)$で割ることによって，テキスト d における単語 i と単語 j の相関値 $pp(d,i,j)$を求めた．以下にその式を示す．

$$pp(d,i,j)=\left(\sum(1/Log((da(i,j)[k])+\beta))\right)/betweenCount(i,j)$$

（式　4.2.3）

そして，全てのテキストに対して，$pp(d,i,j)$を算出，配列 $pp(d,i,j)[]$に代入し，単語の出現位置による全テキストによる単語 i と j の相関値 $prel(i,j)$を以下の式で求めた．

$$prel(i,j)=\sum pp(d,i,j)[n]/df(i,j)$$

（式　4.2.4）

４．３　関連語解析の評価方法

そして，最後に，単語の出現頻度から求めた相関値 $frel(i,j)$と単語の出現位置から求めた $prel(i,j)$，単語 i と単語 j が出現するテキスト数 $dc(i,j)$，単語 i が出現するテキスト数 $dc(i)$より，単語 i と単語 j の関連語の関係値 $rel(i,j)$を求めた．但し，γ，δは定数．

$$Rel(i,j)=(F\,Rel(i,j)+\chi)*(PRl(i,j)+\delta)*(dc(i,j)/dc(i))$$

（式　4.3.1）

以上の式を用いて関連語解析を行う．

５．自動構築においての前処理

５．１　辞書の意味文章への主語挿入

辞書の意味文章への主語挿入を行う．辞書の意味文章には一般的には，主語（つまり見出し語）は省略されている．そこで以下のアルゴリズムを用いて，辞書の意味文章へ主語を挿入した．辞書の意味文章は，見出し語を説明している文の集合であり，係助詞は，『は，には』しか基本的には存在しないであろう，という理論に基づき，係助詞リストを，『は，には』のみとし，辞書の意味文章を一文ずつ形態素解析し，係助詞リストが存在しなければ，『見出し語＋は，』をその文の先頭に追加する処理を行った．そして，例外処理を追加した．例外処理は，文が，例外リスト『Ａを（係助詞，を）Ｂと呼ぶ』等，にマッチしたら，この文は主語が存在するものと見なす処理である．また，時勢や数量を表し，主語とはなりえない，且つ，その文中では重要な意味を持つ単語ではない単語，『現在は，現在には，多くは』等も，この時点で消去した．

５．２　表記の揺れによる同義語の自動生成

次に日本語特有の表記の揺れに対する対策を述べる．表記の揺れとは，例えば，『ＱＣ』（品質管理のこと）は，『Ｑ．Ｃ．』や，『QC』の様に，同じ意味を示すが，表記方法が異なることを示す．本研究では，辞書の見出し語に対する表記の揺れの単語を自動生成し，見出し語の同義語として，同義語テーブルに挿入した．この処理においては，アルファベットの全角半角，アルファベットの大文字小文字，記号の全角半角，カタカナ問題に対応する．

５．３　検索サイトを用いた WEB テキストの所得

検索キーワードの基は，辞書の見出し語（もしくは同義語）とし，まず始めに，『見出し語＋とは』を検索キーワードとして検索する．もし結果リスト数が検索リトライ最大制限テキスト数より多ければ，『見出し語＋とは』と『建築』という二つの単語を検索キーワードとして検索する．もし，検索結果が検索リトライ最小制限ファイル数より少なければ，『見出し語＋は』で検索をし，WEB テキストを取得する．

６．辞書を用いたオントロジーの自動構築

本章では，辞書のみを用いてシソーラス体系を構築する．辞書には，見出し語，見出し語の意味文章が

在る．一見出し語に対して，意味文章は平均約３文存在する．辞書においての意味文章は，Web のテキストファイルに比べて，比較的テキスト信頼度が高いと考えられる．

また，本章において，LOG 関数を使用する場合，底は常に e とする．

６．１　見出し語解析（広義語，狭義語の分析）

まず，見出し語解析を行い，広義語と狭義語の分析を行う．見出し語解析とは，辞書の見出し語を形態素解析ソフト『茶筅』にかけ，最小単位の形態素に分解する．そして，『最終位置に位置する形態素は，最終の一つ前に位置する形態素＋最終位置に位置する形態素の広義語である』（狭義語）というルールを適用し，狭義語，広義語の分析をする．

例）土木用語　＝　土木＋用語

⇒『用語』は，『土木用語』の広義語である．

⇒『土木用語』は，『用語』の狭義語である．

一つの関係を取得する毎に，一ポイントとし，関係毎にポイントを加算していった．その結果，関係ポイント数が『1』のものが全体の約 93%（15350 関係）を占めており，また関係ポイント数が『10 以上』のものは全体の 1%未満（13 関係）しかない．それぞれの関係を見てみると，『10 以上』の関係ポイント数を持つものは，例えば，『水制』と『制』（16 ポイント）等，殆ど一単語と認識してよいものが多い．また，『1』の関係ポイント数を持つものは，比較的間違ってはいないが，信頼度が低い．以下に，見出し語解析結果の一部を示す．（表 6-1）

【表 6-1】　関係カウント数（一部）

狭義語	広義語	関係カウント数
支承	支	16
保工	保	12
交通量	量	10
アーク溶接	溶接	6
剪断試験	試験	5
リフト工法	工法	2

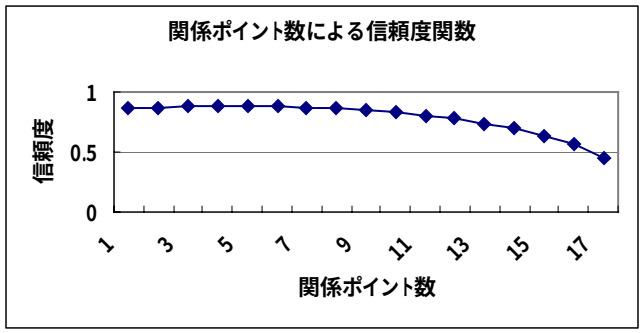
よって，見出し語解析の結果においては，ポイント数が『3～9』くらいの関係に対して妥当性があり，信頼できるものと予測できる．従って，以下の信頼度関数を用いて関係ポイント数に対する信頼度 RP(point)を求めた．但し，point は関係ポイント数とする．

$$PC(point) = (Log((-(point - \alpha)\text{^}2) + \beta)\text{^}2)/\gamma$$

(式　8.1.1)

但し，本研究において，α（4.5），β（200），γ（30）とした．以下に，関係ポイント数による信頼度関数のグラフを示す．（グラフ 6-2）

【グラフ 6-2】　関係ポイント数による信頼度関数



そして，この信頼度関数と関係ポイント数を基に，見出し語解析結果による広義語狭義語関係値 *EntryAnalyzeRe(i,j)*を以下の式を用いて算出する．但し，関係においての i は狭義語を，j は広義語を示す．（ただし，本研究において，α＝1000，β=6）

$$EntryAnalyzeRel(i,j) = RC(point) * Log(point * \alpha)/\beta$$

(式　8.1.2)

以下に見出し語解析の結果を示す．（表 6-3）

【表 6-3】　見出し語解析結果（一部）

狭義語	広義語	評価値
突合せ溶接	溶接	0.7334
流量曲線	曲線	0.6956
候性鋼材	鋼材	0.6411
バンカー線	線	0.5506

６．２　辞書の意味文章に対する評価

次に，ルール解析を行う前に，辞書のそれぞれの意味文章に対して，意味文章に対する評価値を計算する．評価値の計算方法は，それぞれの意味文章中に，その意味文章が対応する見出し語以外の辞書に登録されている見出し語の数（見出し語以外の辞書単語）を計算し，この数を基に，意味文章の評価値を計算する．

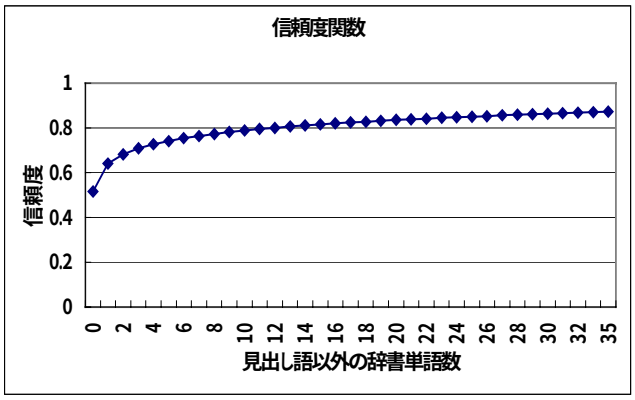
一見出し語の意味文章に対して，見出し語以外の

辞書単語は平均 6.7835 個含まれており，全体の分散は 32.2971 個となった．次に，この見出し語以外の辞書単語数を基に，以下の式を用いて，意味文章に対する信頼度を算出した．以下に式と信頼度関数のグラフ（グラフ 6-4）を示す．

$$reliability(d) = Log(entryCount * N + \alpha) + \beta$$

（式　6.2.1）

ただし，　*EntryCount* は辞書単語の頻度数，*Relaiability* はテキストの信頼度を示す．α，β，N は定数.



【グラフ 6-4】　信頼度関数

６．３　ルール解析

ルール解析を辞書の意味文章に対して実行し，解析する．辞書の意味文章を解析することにより，それぞれのルールに対して，以下の式を用いて単語と単語の関係の信頼度を計算し，シソーラスを構築した.

$$relation(i, j) = reliability(d) * point + relation(i, j)$$

（式　6.3.1）

----- 変数の説明
Reliability(d) ： テキストＤの信頼度
Relation(i,j) ： 単語Ｉと単語Ｊの関係値
Point ： 一ルールに対してのポイント

辞書の意味文章のルール解析を行った結果，2930 個ものツリーを構築することができ，見出し語解析では得られなかった結果を得ることができた．しかし，個々のツリー自体は，ツリーの階層も低いものが多く，単独ツリーのものが目立つ．また，これとは反対に，膨大な数のノード（単語）を持つツリーもいくつか存在する．全体の 85%を占めるノード数が２～４個のツリーは 2481 個存在し，ノイズも含まれているが，比較的，単語と単語の関係（広義語と狭義語の関係）はうまく取得できている．以下に，以上の結果を示す（表 6-5）

【表 6-5】ルール解析結果

狭義語	広義語	関係値
オールクラッシュコスト	費用	0.891107251
コンデンサー	装置	0.895792077
花崗岩	深成岩	0.837719464
片開き分岐器	分岐器	0.719852522
岩盤力学	学問分野	0.730402841

ルール解析の結果，比較的正当性の高い関係を得られることはできた．しかし，階層が 1～3 階層のツリーではシソーラスとは呼ぶことはできない．これらの階層の低いツリー同士に関係を持たす必要があり，方法としては，より多くのファイルを解析することが考えられる．解析対象ファイルは，辞書データではない土木に関する文章が望ましい．従って本研究では，検索サイトを用いて土木に関するテキストを取得し，これらのテキストに対してルール解析処理を適応することを試みた．これについては，次章で述べる.

６．４　関連語に関する単語選択

次に関連語解析を行う．関連語解析には，解析対象ファイルと解析対象単語が必要である．解析対象ファイルは，辞書の意味文章は全て土木に関連した文章であり，かつ信頼度の高い文章なので，辞書の全意味文章とする．一方, 解析対象単語においては，全単語の総当りの関係を求めることは，計算時間が膨大になり不効率である．理由は以下の二点である.

一点目は，全く関係のない単語（見出し語）同士の関係も存在すること，二点目は，解析対象ファイルの文章が，一ファイルあたり平均約 3 文と極度に少ないことである．関連語解析は一ファイル中において単語と単語の関係度を計算し，それを全ファイルで平均する．従って，一ファイルあたりの文の数が少ないと，単語と単語が同ファイル中に出現する確率が減り，全ての見出し語同士の関係度を算出す

ることが難しい．以上の理由により，本工程における，関連語解析対象単語の選択方法は，以下の二つの基準を用いて選択した．

(ア)見出し語の意味文章中に，その見出し語以外の見出し語の数が多い見出し語
(イ)ルール解析結果，抽出関係が多い見出し語

まず，①については，本章の 7.2（辞書の意味文章に対する評価）で算出した信頼度関数（以下，本章において，この関数を『信頼度関数α』とする.）を用いる．この信頼度関数は，見出し語の意味文章中に含まれる見出し語以外の見出し語数を基とした関数であり，意味文章中に見出し語以外の見出し語の数に比例して対数関数的に，増加する関数である．

次に，②の条件は，ルール解析結果，抽出された関係が多い見出し語を優先的に選択するという条件である．

6．5　関連語解析

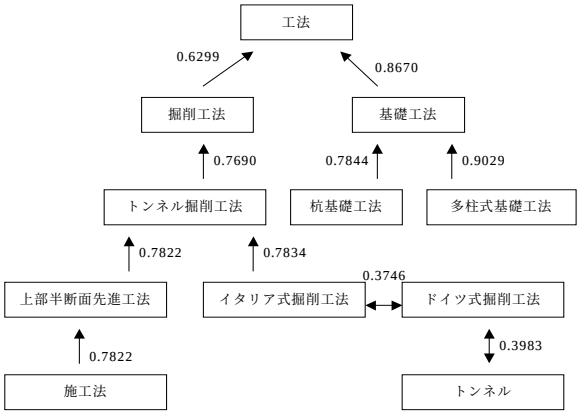
本工程では，関連語解析について述べる．以下に，関連語解析を行った結果を示す．（表 6-6）

【表 6-6】　関連語解析の結果

関連語 I	関連語 j	関係値
ロープ	ケーブル	0.961080373
レール	枕木	0.875993363
噴出岩	火山岩	0.676010963
開水路	放流管	0.663696852

6．6　総合結果

見出し語解析，ルール解析，関連語解析のそれぞれの結果を統合した結果を示す．（図 6-7）



【図 6-7】　辞書を用いたオントロジー自動構築の結果（一部）

6．7　考察

本章において，辞書のみを対象に，見出し語解析，ルール解析，関連語解析を行い，シソーラス構築を試みた．ルール解析における同義語予測の結果は，ノイズが目立ち，実用に絶えられるものではなかった．ルールの適応能力の低さが原因の一つだと考えられる．

見出し語解析における広義語狭義語の予測に関しては，最初はノイズが多かったが，信頼度関数を用いることにより，精度の向上が見られた．一方，ルール解析における広義語狭義語の予測においても，ノイズは含まれるものの，抽出された関係の信頼性は比較的精度は良い．

関連語解析における関連語の予測は，解析対象ファイルが辞書の意味文章と，位置テキストあたりの文が少ないにもかかわらず，良い結果が得ることができた．

7．WEB テキストを用いたオントロジーの自動構築

7．1　WEB テキストの評価方法

WEB テキストは，ファイル作成者は様々であり，内容の正当性も様々である．また，今回は，土木辞書の見出し語をキーワードとして検索サイトから検索し，その結果ファイルを取得している．そのため，土木分野における単語の意味と一般的に認識されている単語の意味が違う場合，一般的な意味で認識されている意味で使われている WEB テキストがヒットしている可能性が高い．従って，それぞれの WEB テキストに対して評価値を計算し，考慮する必要がある．WEB テキストの評価値は，テキスト信頼度，テキスト土木度の二つの値から求める．テキスト信頼度とは，その WEB テキストのどのくらい信頼できるかという度合いを示し，テキスト土木度は，そのテキストがどれほど土木に関連しているのかという度合いを示す．以下に評価パラメターに用いたチェック項目を示す．（表 7-1）

以下のパラメターを，全 WEB テキストに対して求め，それぞれの WEB テキストの評価パラメータとする．

【表 7-1】 WEB テキスト評価項目表

チェックする個所	チェック項目	度合い
URL	yougo・yogojisho・lib・library　の有無	信頼度
<title>タグ	辞書・用語・建築・とは　の有無	信頼度
<title>タグ	そのWEBテキスト取得時に用いた見出し語キーワードの有無	土木度
<b>タグ	その WEB テキスト取得時に用いた見出し語キーワードの有無	土木度
<h>タグ	その WEB テキスト取得時に用いた見出し語キーワードの有無	土木度
<font>タグ	その WEB テキスト取得時に用いた見出し語キーワードの有無	土木度
テキスト全体	辞書の意味文章中に見出し語以外の見出し語が 10 以上存在する見出し語のうち，検索結果が 15 件以上もつ見出し語の総数	土木度
テキスト全体	辞書の意味文章中に見出し語以外の見出し語が 10 以上存在する見出し語のうち，検索結果が 15 件以上もつ見出し語の種類	土木度

７．２　ルール解析用ファイル選択判別器自動構築

検索サイトより取得した WEB テキストに対しても，ルール解析処理を行う．しかし，WEB テキストの量は膨大であり，全てを解析することは実時間内で不可能であり，また不効率，無意味である．従って，本研究では，WEB テキストの土木度が高く，かつ，信頼度の高い WEB テキストのみに対してルール解析の対象ファイルとする．前コラムにおいて，WEB テキストに対して評価パラメターを計算した．しかし，どのパラメターがどれだけの値を満たせば土木的であるか，信頼できるかは，一般的には人間の主観に依存するものであり，誰もが認める完璧な判別ができるとは言えない．

従って，本研究では，WEB テキストに対して土木的であるか土木的でないか，信頼できるか信頼できないかという判別を下す判別器を，Ensemble Learning(Bagging 手法と Boosting 手法)を用いて自動生成することを試みた．Bagging 手法とは，同データベースから複数回のサンプリングを繰返し，それぞれの結果の多数決で最終的な結果を決定する．Boosting 手法とは，基本的には Bagging 手法と同じであるが，データのサンプリングを間違ったデータに対して集中的に行う．以下に評価関数自動生成について述べていく．

まず，最終的な評価関数のイメージから説明する．最終的な評価は，以下の概念図の様に，WEB テキストを各判別器にかけ，その結果，各判別器（*f1(x)*～*fn(x)*）からの重み付き投票（α1～αn）による重み付き多数決となる．各判別器（*f1(x)*～*fn(x)*）は，前コラムで述べた『WEB テキスト評価項目』の一つ一つである．α1～αn は，重み付き投票であり，各判別器に重み *g(n)*を付け，各判別器から出力された結果に対して，その重み *g(n)*を掛けたものとする．

つまり，もし入力値（WEB テキスト）が関数 *fn(x)* の条件を満たすならば，*fn(x)***g(n)*の値を『土木的である投票箱』へ，もし関数 *fn(x)*の条件を満たさないならば，*fn(x)***g(n)*の値を『土木的でない投票箱』へ投票するというフローである．

次に，生成する関数一覧を以下に示す．（表 7-2）

【表 7-2】　生成する関数と重み一覧

関数	重み	チェックする個所	チェック項目
f1(x)	g1	WEB テキストの URL	yougo・yogo・jisho・lib・library の有無
f2(x)	g2	<title>タグ	辞書・用語・建築・とはの有無
f3(x)	g3	<title>タグ	その WEB テキスト取得時に用いた見出し語キーワードの有無
f4(x)	g4	<b>タグ	その WEB テキスト取得時に用いた見出し語キーワードの有無
f5(x)	g5	<h>タグ	その WEB テキスト取得時に用いた見出し語キーワードの有無
f6(x)	g6	<font>タグ	その WEB テキスト取得時に用いた見出し語キーワードの有無
f6(x)	g7	テキスト全体	辞書の意味文章中に見出し語以外の出し語が 10 以上存在する見出し語 のうち，検索結果が 15 件以上もつ見出し語の総数

f7(x)	G8	テキスト全体	辞書の意味文章中に見出し語以外の見出し語が 10 以上存在する見出し語のうち，検索結果が 15 件以上もつ見出し語の種類数

上の表の関数 fn とそれぞれの関数 gn の重みを生成する．

７．２．１　関数 fn の生成

関数 fn は，次の二つのタイプに分類できる．

Ａ　：　関数の出力値が『真｜偽』のもの

Ｂ　：　関数の出力値が『数値』のもの

f1(x)～f6(x)は A タイプ，f7(x)，f8(x)は B のタイプに属する．まず，A タイプについてから述べる．A タイプにおいての関数の場合，二値しか取らないので，そのままの形式の式（真⇒１，偽⇒０）を用いることとする．次に B タイプについて述べていく．B タイプも入力値（WEB テキスト）対して，出力値は『真｜偽』のいずれかにしなければならない．しかし，B タイプにおいては，評価値が両方とも正の整数（土木辞書の見出し語総数，土木辞書の見出し語の種類数）である．従って，しきい値の様なものを設定しなければならない．しかし，しきい値を一意に決定することは困難であり，また正当性や信頼性にも不安が残る．

従って，本研究では，これらを解決するための判別器アルゴリズムを Boosting 手法を用いて導き出す．具体的には，『ある WEB テキスト内に，土木辞書に登録されている見出し語が N 個見つかった．この WEB テキストは，土木的であるか，もしくは土木的でないか』ということを判別する判別器である．

以下に本研究で行った Boosting 手法を用いた判定機の生成方法について述べていく．

学習用の例題テキストには，検索サイトより取得した WEB テキストを用いた．土木的である例題 WEB テキストは土木分野の用語をキーワードとして取得した WEB テキストを，土木的でない例題 WEB テキストは IT 分野の用語をキーワードとして取得した WEB テキストを用いた．まず，上記の表（表 7-2）にある全てのパラメターを求めた．そして，これらのパラメターをもつ例題を用いて，Boosting 手法により，判別器を生成する．次に，この Boosting 手法を用いて生成された判別器の多数決により，WEB テキストが土木的であるかどうかを判別する一つの判別器と見なす．

つまり，この一つの判別器が，B タイプの判別器となる．しかし，Boosting 手法により生成した結果はあまり精度の良いものではなかった．ノイズを多く学習することにより，精度の向上を求めていたが，逆にノイズのみを学習し，判別器がノイズを判別しきれなかったことが原因だと考えられる．従って，本研究では，Bagging 手法を用いることとした．

７．２．２　重み gn の生成

次に，重み *g(n)*の生成方法について述べる．重み *g(n)*は，例題をランダムにサンプリングする Bagging 手法を用いて生成した．

７．３　ルール解析

次にルール解析を実行する．ルール解析対象ファイルは，前工程で生成したファイル選択関数を用いてファイルを選択する．

７．４　関連語解析における単語とテキスト選択アルゴリズム（関連語解析の前処理）

本工程では，WEB テキストを対象に関連語解析を行う．関連語解析において，第七章でも述べたように，解析対象単語と解析対象ファイルを選択する必要がある．解析対象ファイルは，前工程の WEB テキストを用いたルール解析のテキスト選択手法と同じものを用いる．解析対象単語選択方法においては，第七章でも述べたように，全単語の総当りの関係を求める必要はない．なぜなら，全く関係のない単語（見出し語）同士の関係も存在するからだ．しかし，関係を調べる必要がある単語と必要の無い単語の判断を人間が行うのは，非効率的であり，またその人の主観が入ってしまう恐れ大いにある．従って，本研究では，ルール解析によって構築された広義語狭義語の関係ツリーを基に，解析対象単語を選択した．

７．５　関連語解析

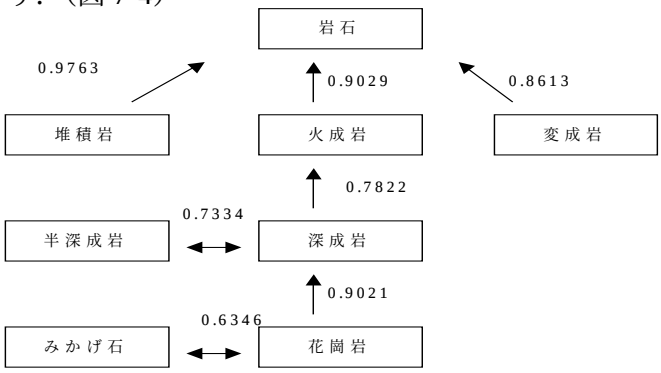
以下に結果を示す．

【表 7-3】　関連後解析結果

関連語 I	関連語 j	関係値
逆巻工法	順巻き工法	0.95308
ブロック張り工	風化	0.92815
スタッド溶接	プラズマ溶接	0.85517
モルタル	セメント	0.53923

7．6　総合結果

以下にルール解析と関連語解析の結果の一部を示す．（図 7-4）



【図 7-4】　結果図

7．7　考察

検索サイトを用いて，ルール解析，関連語解析を行った結果，辞書のみよりも良い結果を得ることができた.WEB テキストの方が辞書よりも信頼度が低いにもかかわらず，良い結果が得られたのは，Bagging 手法を用いたファイル選択が精度がよかったからだと考えられる．しかし，一方，ルール解析，関連語解析での精度は思ったより低かった.

8．　おわりに

本研究では，辞書と検索サイトを用いてオントロジー，シソーラスの構築を試みた．信頼性の高い辞書と，信頼性は低いが数の多い WEB テキストを用いて解析処理をすることにより，信頼のできる関係を抽出することができた.

しかし，ノイズが多く存在し，全ての関係を高い信頼性で構築することはできなかった．これは，ルール解析，関連語解析の精度の低さと，WEB テキスト自体の信頼度の低さ，また WEB テキストのゴミ取り処理（HTML タグなど）が十分に処理しきれなったことが原因だと考えられる.

今後の課題としては，膨大な量のデータに対して，ノイズを除去しきれなかったことがあげられる．これらを改善していけば，もっと少ない量のデータで，信頼度の高いのオントロジーを，自動構築できると考えられる.

参考文献

[1] 山田篤, 安達文夫, 海田茂, 今門政記, 河合正樹, 小町祐史：博物館情報の知的横断検索のためのフレームワーク，画像電子学会 VMA 研究会博物館・美術館 DTD-SG

[2] 国分芳宏：シソーラスとオントロジー，株式会社言語工学研究所　(2002 年 9 月 9 日)

[3] 伊藤英毅:オントロジーを利用した知識の共有／再利用，UNISYS TECHNOLOGY REVIEW 第 64 号(2000 年 2 月)

[4] 西村広光，小林誠，丸山稔，中野泰明：多方向特徴抽出 HMM と Bagging による多数決を用いた文字認識，電子情報通信学会論文誌，D-Ⅱ，Vol.J82-D-Ⅱ，NO.9 ，p1429-1434 (2002 年 9 月)

[5] 石川誠一，久保成毅，古崎晃司，來村徳信，溝口理一郎：タスク・ドメインロールに基づくオントロジー構築ガイドシステムの設計と開発，人工知能学会論文誌，17 巻 5 号 G，pp585-597 (2002 年)

[6] 來村徳信，溝口理一郎：オントロジー工学に基づく機能的知識体系化の枠組み，人工知能学会論文誌「設計知識と管理の高度利用」,論文特集 Vol.1, No1, pp61-72，(2002 年)

[7] 村田昇:Boosting の幾何学的構造と統計的性質，科研費特定領域研究「確率的情報処理への統計力学的アプローチ」(2002 年)

[8] 溝口理一郎：オントロジー工学の試み，1998 年度人工知能学会全国大会（第１２回）AI レクチャ (1998 年)

[9] 馬見塚拓，安倍直樹：集団能動学習-データマイニング・バイオインフォマティックスへの展開-，電子情報通信学会論文誌　D-Ⅱ　Vol.J85-D-Ⅱ NO.5 p717-724 (2002 年 5 月)

[10] 飯田龍，徳永奏浩，乾健太郎，衛藤純司：言い換えエンジン KURA を用いた節内構造および機能語相当レベルの言い換え，第 63 回情報処理学会全国大会　3H-03 (2001 年 3 月)

[11] 松本裕治：自然言語処理におけるシステム混合法の利用，電子情報通信学，D-Ⅱ，Vol.J85-D-Ⅱm,

No.5, pp709-716 (2002 年 5 月)
[12] 太原育夫：選好的仮説集合をもつ知識からのシナリオ計算，電子情報通信，D-Ⅰ，Vol.J85-D-Ⅰ，No.8 pp776-783 (2002 年 8 月)
[13] 笠晃一，弘中大介，天野幹郎，横田将生：心臓意味論を組み込んだ HPSG による意味解析システ，電子情報通信学会，D-Ⅱ，Vol.J85-D-Ⅱ，No.3 pp475-482 (2002 年 3 月)