知識プロバイダにおけるオントロジ自動獲得

3AF−4

湯川 高志　松澤 和光

NTT(株) コミュニケーション科学研究所

1 はじめに

筆者らは、意図の顕在化とそれに基づく情報の構造化により人間の知識創造とコミュニケーションの能力を拡大させる知識プロバイダを提案している [1]。知識プロバイダの構成要素の中で、情報の集合からユーザの要求意図に沿うような知識を自律的に抽出・構成する部分を「知識オーガナイザ」と名付け提唱する。知識オーガナイザは、要求項目に関連する大量の文書から語の意味や用法の体系 (オントロジ) を自動的に獲得し、これに基づき個々の文書の意味を理解する。そして、理解された大量の情報から知識を抽出・構成する (図 1)。

本稿では、知識オーガナイザ実現のための技術について述べるとともに、その第 1 ステップにあたるオントロジの自動獲得法を提案する。提案する方法は、文書上の語の統計的情報に加えてシソーラス [2] や辞書定義 [3] 等日常語に関する知識を利用することで、連想関係のみの弱構造化オントロジ [4] よりも少し強い構造を持ったオントロジの自動獲得を目指したものである。

2 知識プロバイダの知識抽出・構成部 — 知識オーガナイザ —

知識プロバイダでは、単なる情報の羅列にとどまらない構造化された知識を、広いドメインにわたって提供できる知識抽出・構成システムが必要とされる。情報検索システムやそれを発展させた情報フィルタは取捨選択した生の情報を提供するだけであり、一方、エキスパートシステムは高度な判断を限られたドメインでしか提供しない。知識オーガナイザは、これらの中間を狙いとしており、ユーザの要求に関連する大量の情報から、「見識」といえる程には深くはないが「受け売り」よりも整理された知識 — 「耳学問」レベルの知識 — を抽出・構成し提示するシステムである (図 2)。情報ソースとしては、WWW ページ、グループウェア・パソコン通信・ネットニュースの電子会議記事、そして個人が蓄積した文書等の大量の平文テキストを対象とする。

知識オーガナイザは次の 3 つのステップにより知識を抽出・構成する。

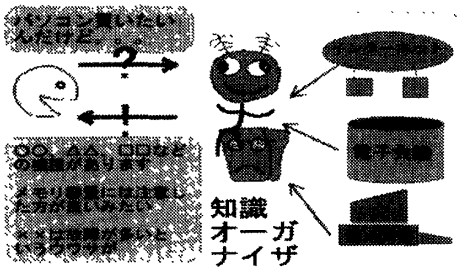

図 1　知識オーガナイザ

(1) オントロジの自動獲得　未知のドメインについて知識を得るには、まずそのドメインにおけるオントロジを獲得することが必要となる。WWW ページや電子会議を対象とした場合、フレーム型言語や一階述語によるトップダウン型のオントロジ構築が現実的でないのは岩爪らが指摘する通りである [5]。知識オーガナイザでは、文書からの知識フレーム抽出への利用を目的とし、要求項目を中心とした有向グラフとしてオントロジを獲得する。

(2) オントロジを利用した知識フレームの抽出　オントロジの獲得によって、要求項目に関してどのような項目が重要かが判断できるようになる。これに基づき文書の内容を知識フレームとして抽出する。MUC[6] で議論されるような文書理解技術を利用する。

(3) 大量の知識フレームからの推論　上で得られた知識フレームから知識を体系化しユーザに提示する。文書そのものが常に正しい内容であるとは限らず、また知識フレーム抽出においても誤りは避けられないため、上で得られる知識フレームには誤りや矛盾が含まれる。不完全で大量の知識から確らしい結論を導き出す「質より量」の推論技術 [7] を利用して、これらフレームから知識を体系化する。

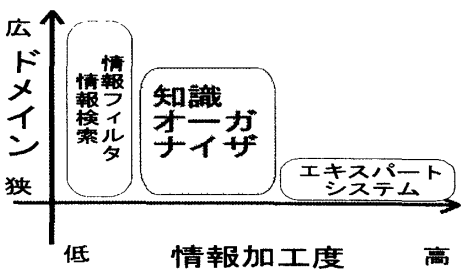

図 2　知識オーガナイザが対象とする領域

Automatic Acquisition of Ontology for the Knowledge Provider
Takashi YUKAWA and Kazumitsu MATSUZAWA
NTT Communication Science Laboratories
1-1 Hikari-no-oka Yokosuka-shi KANAGAWA 239 Japan

3　オントロジ自動獲得法の提案

文書そのものからは語に関する統計的情報を得ることができる。また、筆者らは日常語に関する構成的な知識ベースを既に有している[2, 3]。これらを利用し、要求項目に関連の深い語の抽出と、共起頻度・日常語知識に基づいた語間のリンク形成とによりオントロジを自動獲得する方法を提案する。具体的には以下の手順により有向グラフの形式でオントロジを構築する。

(1) 関連語抽出　文書から要求項目に対する関連語を抽出し、これらの語に要求項目から「関連」リンクを張ったフラットなグラフを初期グラフとする。関連語とは要求項目との共起頻度の高い語である。

(2) リンク形成　初期グラフに含まれるすべての語の組について、以下を調べそれぞれ有向リンクを張る。

共起関係:　語Aと語Bが同時に出現する文・段落・文書の割合が語Bのみの出現に比べて大きい場合、語Aから語Bに向かって「共起」リンクを張る。

類似関係:　2語のシソーラス上の距離がある値より短い場合および辞書定義に基づく類似度[3]が高い場合に双方向に「類似」リンクを張る。

(3) グラフ変換　以下の変換ルールに基づいてグラフ変形を行う。

従属化変換:　図3(a)に示すように語Bから語Cに対して「共起」リンクがある場合、語Aからの語Cへの「関連」リンクを切り語Bから語Cへのリンクを張る。

クラスタ化変換:　図3(b)に示すように、語Bと語Cに双方向の「共起」リンクまたは「類似」リンクがある場合、両者の直上に仮想的なノードDを設ける。仮想的なノードにはシソーラスや辞書情報に基づいてラベルを付与する。

また、変換時に品詞や日常語知識から関係が推定できる場合には、それに基づきリンクにラベルを付与する。

4　オントロジ獲得例

PC周辺機器であるSCSIカードに関するオントロジ獲得を題材としたグラフ変換の例を図4に示す。

図4(a)では、初期に「カード」、「A社」(実際には実在の会社名)が得られており、「カード」から「A社」へ共起リンクが検出されている。このため変形ルール1によって「A社」は「カード」に従属する。図4(b)では、「高い」、「安い」という語に対し双方向に「類似」リンクが検出されている。このため、新たなノードが設けられ、両ノードはそれに従属する形になる。新たなノードは、シソーラス情報の利用により「値段」のラベルが付与される。

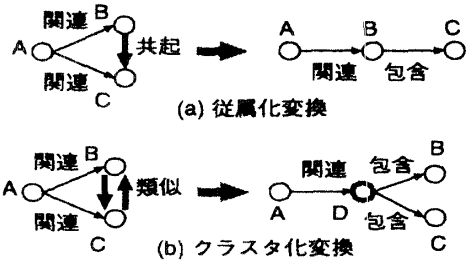

図3　グラフ変換ルール

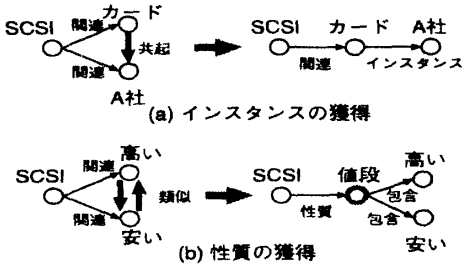

図4　オントロジ獲得例

5　まとめと今後の課題

知識プロバイダの構成要素であり、大量の情報集合から知識を構成する知識オーガナイザを提唱した。知識オーガナイザの実現に必要なオントロジ自動獲得、知識フレーム抽出、大量で不完全な知識に基づく推論について述べ、オントロジの自動獲得法について提案した。本獲得法は、語に関するあらゆる情報を利用することにより、弱構造化オントロジよりも強い構造を持ったオントロジを構築する。また、実際の電子会議の記事に本方法を適用した例を示した。

今後は、様々な要求項目に対してのオントロジ自動獲得実験を行なうとともに、自動獲得されたオントロジの品質評価を行なう予定である。

参考文献

[1] 八巻他「知識プロバイダ構想の提案」, 第55回情処全大,3AF-2,1997.

[2] 池原他「日英機械翻訳のための意味解析辞書」, 情処自然言語処理研究会,84-13,1991.

[3] 笠原他「国語辞書を利用した日常語の類似性判別」, 情処論,Vol.38, 1997.

[4] 岩爪他「インターネットからの情報収集・分類・統合化のためのオントロジー獲得」, 第10回AI全大,18-03,1996.

[5] 岩爪他「オントロジーを用いた情報の自動収集と分類へのアプローチ」, 第9回AI全大,15-06,1995.

[6] "Message Understanding Conference (MUC-5)," 1993.

[7] 松沢他「アバウト推論における記号とパターンの統合」, 第10回AI全大,S4-04,1996.