大規模知識ベースマシン実験機の開発（４）

－単一化エンジンの評価－

3B-4

小黒雅己　森田幸伯　伊藤英則　横田治夫　高橋正寿　白瀬勝次
NTT電気通信研究所　ICOT　富士通　日科技研

1．はじめに

関係モデルを拡張し、属性値が変数を含むある種の構造体である項の集合を、二次記憶に格納した関係型知識ベースマシンの開発を行っている。本マシンの実現には、二次記憶からの項の検索を集合対集合の演算で行うためのRBU演算(Retrieval-By-Unification)を高速に実行する単一化エンジン[1] が必須要素となる。その構成方式については、文献[2] で示した。

本稿では、この構成方式について、ソフトウェア・シミュレータによる評価を行う。なお、本評価では、一回のデータに対する操作（read/write）に要する時間を１クロックとする。

2．全体と各部の処理概要

本単一化エンジンは、ソータ（SU）、ペア生成部（PGU）、および単一化処理部（UNU）の3部から構成される[1] 。これらがパイプラインで動作する。これら各部と全体の処理クロック数／入力特性を図１に示す。

全体の処理は、UNUの処理にほぼ一致している。これは、SUでの項の整列やペアの生成は、UNUでの単一化の処理より軽いからである。一方、UNUの入力量は、PGUの出力量である。PGUの出力量は、PGUの単一化失敗項組の削減機能（以降、候補削減能力と呼ぶ）で定まる。

以上より、エンジンの性能は、PGUの候補削減能力とUNUの処理能力に左右される。以下、各部の性能を評価する。

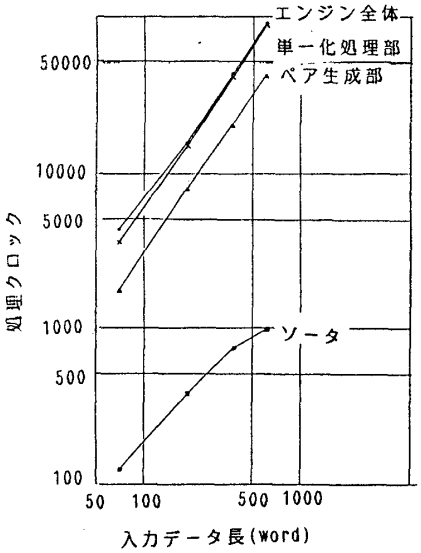

図１　全体と各部の処理関係

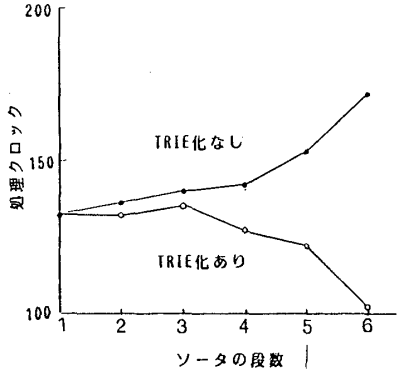

図２　TRIE化の効果

3．ソータ

ソータは、2way-merge-sort を可変長のデータ（項の系列化表現）に対し行う。この処理は、固定長と同様に入力データ長に比例する。また、ソータは、項集合のTRIE化処理を行う[1] 。このことにより、構造が類似した項に対しては、流すデータ量を削減できる。TRIE化の効果例を図２に示す。後段に行くに従い、TRIE化が進み、処理負荷が小さくなる。

4．ペア生成部

PGUの候補削減能力を評価する。表１に、PGU出力項組数／単一化項組数特性を示す。この能力を、木構造一定で中間・葉ノードの関数子の種類の数が変化する場合と、木構造自身が変化する場合に分けて評価する。表１では、木構造が複雑になるか、または木構造が簡単であっても関数子の種類数が多く存在する程、単一化項組数に対してPGUの出力項組数が多くなる。従って、この様な場合に、PGUの候補削減能力が低下し、エンジン全体の性能が低下する。

次に、項のオーダリングの方式の違い(Left-most方式とOuter-most方式[1])による候補削減能力を評価する。PGUは、記号列で表した項間の最初の変数までの一致性をチェックする。項の変数は、葉ノードにしか現れない。このため、変数より上位のノードにある関数子の、種類数が多い場合は、Outer-most方式の候補削減能力が優れ、変数より左のノードにある関数子の、種類数が多い場合は、Left-most 方式の候補削減能力が優れる。

5．単一化処理部

種類数が変化するノード	関数子の種類数		
	1	2	3
根	1.08	1.08	1.08
中間	1.08	1.86	2.63
葉	1.08	1.69	2.73

(a) 木構造を一定にした場合

木構造は、深さ２
根ノードがｎ引数関数１種
葉ノードが定数３種と変数

n	1	2	3	4
	1.00	1.35	2.08	3.13

(b) 木構造を変化させた場合

表１　ペア生成部の性能
（PGU出力項組数／単一化項組数）

Development of an Experimental Very Large Knowledge Base Machine (4)　-Evaluation of Unification Engine-
M.Oguro (NTT Electrical Communications Laboratories)　Y.Morita　H.Itoh (ICOT)　H.Yokota (Fujitsu Ltd.)，
S.Takahashi　K.Siranami (The Institute of JUSE)

5.1 パイプライン処理

UNUの各エレメントの処理は、パイプライン化を行う。このパイプライン化の効果を、UNION-FINDメモリを用いて、一組の項の単一化の高速化を行なった文献[3] と処理速度を比較して、表2に示す。一組の項の単一化では、本方式は処理が遅くなるが、複数組の項の単一化には、パイプライン化が効果的である。

表2 パイプライン化の効果

t1=f(a1,a2,...,ak)とt2=f(X1,X2,...,Xk)の単一化

	単一化対象項数				
	パイプライン				文献3
k	1：1	5：5	20：20	50：50	1：1
1	18	9	8	6	10
2	21	12	9	8	13
10	56	37	27	26	37
25	131	76	56	−	82

単位：クロック／ 項組

5.2 単一化処理方式の考察

文献[2] で述べた単一化処理方式は、置換随時適用方式（SAF）であり、以下の特徴がある。

①単一化セルには、単一化結合対象属性の項が流れ、変数の置換(substitution)を随時求める。

②置換えセルには、入力属性タプルが流れ、これに、置換の適用を随時行なう。

SAFでは、②により、次の問題が起きる。

後段に行くに従い置換えセルの処理量が増える。さらに、結果的には単一化が失敗する項組（失敗組）の入力属性タプルも置換えセルを流れ、これに無駄な置換の適用を行っている。

この問題点を改善するため、図3に示すmgu適用方式（MAF）を提案した。MAFには、以下の特徴がある。

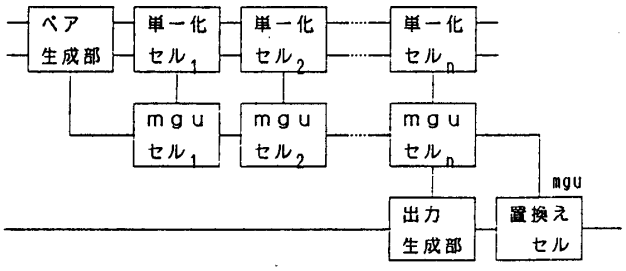

図3 mgu適用方式

1)結果的に失敗する組の入力属性タプルへの置換の適用をなくすため、mguを求めた後、単一化成功項組の入力属性タプルにのみmguを適用する。

2)mguを求めるために、単一化セルで求まった置換を、集合にしてmguセルに流す必要がある。

この2方式の比較を行う。評価実験として、UNU入力項組数一定で、その中に含まれる失敗組数を変え、単一化処理クロック／UNU出力データ長特性を測定した結果を図4に示す。図4の破線は、失敗組が無いときの特性である。図4から以下が言える。

Ⅰ．失敗項が無ければ、処理クロックは、出力データ長による。

Ⅱ．交点Pの出力データ長以上では、2方式の処理クロックは、ほぼ一致する。

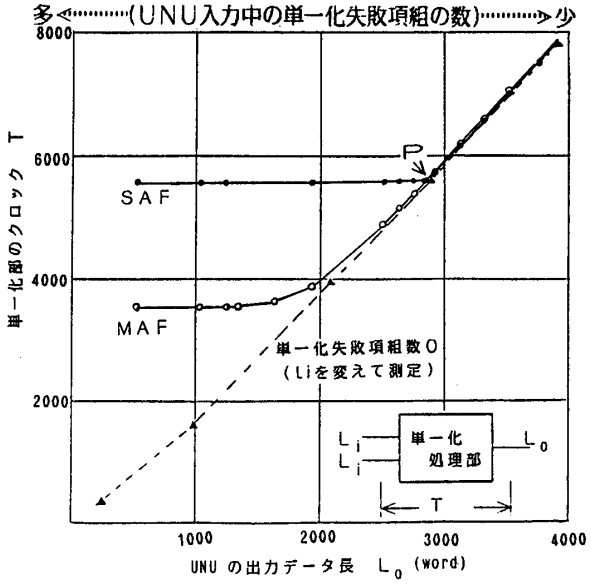

図4 失敗による処理クロックの違い （L_i =3265一定）

Ⅲ．失敗組が多い時、両方式とも、出力データ長によらず処理クロックが一定となり、しかも、MAFのそれはSAFより小さい。

これらより、以下が考察される。Ⅰは、条件より明らかである。Ⅱは、失敗組が少ないとき、UNUの出力量（L_0 ）が、他のどのセルを流れる量よりも多くなり、処理時間がL_0 で決まるからである。また、失敗組が多いとき、SAFでは、L_0 よりも、最終段の置換えセルの入力量、MAFでは、L_0 よりも、最終段のmguセルから出るmguの量の方が多くなり、処理時間がそれぞれの量で決まる。Ⅲは、これらの量が、UNUの入力（L_i ）一定の時、ほぼ一定の量になるためである。

以上より、置換集合の量が入力属性タプルの量より小さいことを前提にすると、UNU入力に失敗項が多く含まれる場合は、UNUの処理は、置換随時適用方式よりもmgu適用方式が優れる。

6．おわりに

単一化エンジン性能は、UNUの性能で決まる。PGUでは、項構造が複雑になると候補削減機能が低下し、単一化失敗項組が多く残るため、UNUに、置換随時適用方式を採用すると性能が低下し、エンジン全体の性能低下につながる。このため、UNUに、mgu適用方式を採用することで、単一化失敗項組が多く含まれるデータに対するPGUの候補削減能力低下を補償できた。今後の課題には、UNUの処理方式として、スイッチングネットワーク[1] の評価を行い、シミュレーション解析を反映して、単一化エンジンの詳細設計を行う。

参考文献

[1]Morita,Y."Retrieval-BY-Unification on a Relational Knowledge Base Model"12th.VLDB.1986

[2] 酒井他，「単一化エンジンの構成方式」第33回情報処理学会全国大会(1986)

[3] 安福他，「論理型言語の単一化操作のためのハードウェアアルゴリズム」EC84-67