知識ベースマシン *Mu-X*(4)
−項に対するインデックス方式の評価−

仲瀬　明彦、柴山　茂樹　　森田　幸伯
(株)東芝　総合研究所　　ICOT

1.　はじめに

項を基本要素とする知識ベースマシンでは，大量な項にインデックスを付加して単一化を伴なう検索を高速化する技術が重要である．項にインデックスを付加する技法は，Prolog，演繹データベース等においていくつか提案されており，その有効性が確認されている．筆者らは，項に対するインデックス手法としてSSCW(Structual Superimposed Code Word)方式を提案している．SSCW方式の詳細内容，他のインデックス方式との比較は[1] に述べられているが，本稿では，SSCWを様々な特性の項集合に適用した場合の選択性能の変化と，その結果を考慮してSSCW作成時のパラメータの決定方法について述べる．

2.　SSCWの概要

ある項のSSCWによるインデックスを作成する場合は，先ず対象項を木構造に変換する．次に変換された木の各ノードのハッシュ値を計算するが，この際に，親ノードのハッシュ値のビット幅が子ノードのハッシュ値のビット幅の合計より大きくなるように設定する．変数のハッシュ値に対しては，検索されるデータの項内に含まれる変数には全てビット'1' の列を使用し，検索する質問の項内に含まれる変数には全てビット'0' の列を使用する．最後に，親ノードのハッシュ値と横方向に並べられた子ノードのハッシュ値の各ビットごとの論理和をとることによりインデックスを作成する．検索されるデータの項のｉ番目の要素をＤｉ，検索する質問の項をＱとした場合Ｄｉ ∧Ｑ＝ＱとなるＤｉ をＱと単一化検索可能候補とする．SSCWを用いて項" a(b(X),c) " のインデックスを作成する例を第１図に示す．

3.　SSCWの性能評価方法と評価項目

ランダムに発生した 100個の項からなる項集合をいくつか用意し，同一の項集合間で，単一化を含む結合操作を行った．この際に実際に単一化可能な項のペアの数ＲとSSCWによって選択された単一化可能な候補の項のペアの数Ｓをカウントし，失敗率（＝（Ｓ－Ｒ）／Ｓ）を測定することによりSSCWの性能を評価した．評価項目は以下の２点とした

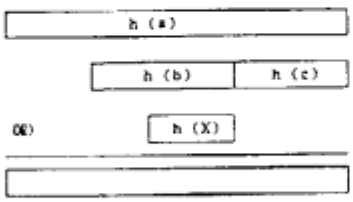
(a) 項集合内に出現する定数の種類が増加した場合にハッシュの衝突がSSCWに与える影響を測定するために，項集合内の定数の種類を40種～ 300種まで変化させて，失敗率の変化を調べる．項集合内の変数の割合いは3%程度に固定する．

(b) 項集合内に出現する変数の割合いが増加した場合に変数のハッシュ値がSSCWに与える影響を測定するために，項集合内の変数の割合いを5%～50% まで変化させて，失敗率の変化を調べる．項集合内の定数の種類は30種程度に固定する

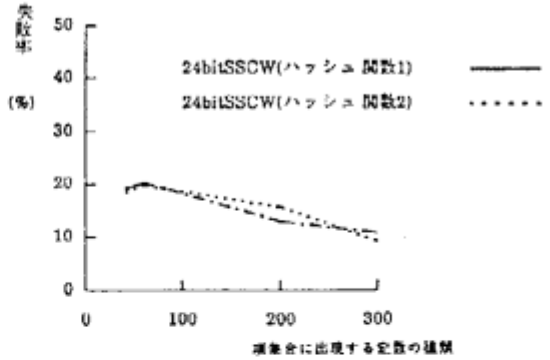
4.　評価結果

4.1 （a)に対する測定結果を第２図(a) に示す．SSCWでは，各定数に対するハッシュ値のレンジを幅広く設定できるため定数の種類が多くなってもハッシュの衝突が生じにくく，項内の定数の種類の大小の影響は少ない．グラフにおいて，定数の種類が少い時に失敗率の多少大きい原因は，定数の種類が少ないために，関数子と形状が類似している項が項集合内に多く現れるためである．

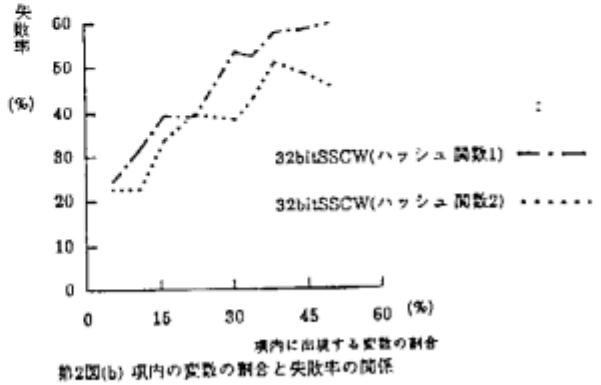
4.2 （b)に対する測定結果を第２図(b) に示す．項内の変数が多くなるにつれて選択性能が悪くなる．



第1図 項" a (b (X), c)" のSSCW



第2図(a) 項内の定数の種類と失敗率の関係



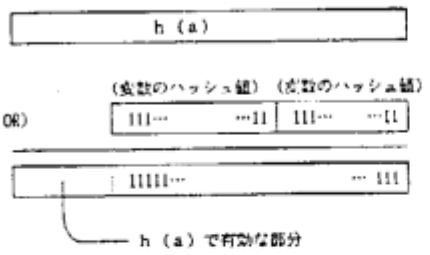
第2図(b) 項内の変数の割合と失敗率の関係

Knowledge Base Machine Mu-X (4)
(Evaluation of Indexing Scheme for Terms)
Akihiko Nakase, Shigeki Shibayama, (Toshiba Corporation)
Yukihiro Morita (ICOT Research Center)

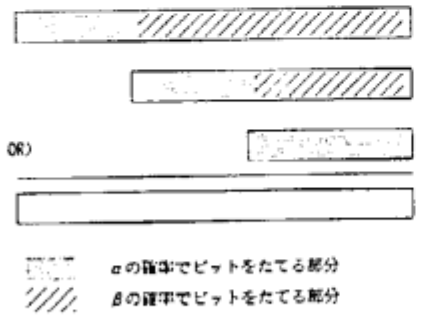
項内の変数が多い時の性能低下の理由を以下に示す．SSCWは定数部分のハッシュ値でビット'1'を立てる確率によって選択性能が左右され，本測定では長さLビットのフィールドにN個のビット'1'の設定されるハッシュ関数を採用している．ハッシュ値でビット'1'の立つ確率（N／L）の最適値についての解析は[1]で述べられている．しかし，あるハッシュ・ビット・フィールドにおいて，一様な確率でビット'1'が立つようなハッシュ関数を用いると，多くの変数を引数にもつ関数子では，第３図に示すように関数子のハッシュ値の情報の大部分が有効に使用できなくなる．このため変数を多く含む項集合にSSCWを適用した場合は，選択性能が低下するのである．

5．　変数の多い項を扱う場合の改良案

以上述べた性能低下を防止するには，ある関数子のハッシュ値の有効な情報をなるべく引数のハッシュ値の影響の及ばないところに集中させる必要がある．そこで，各定数に対するハッシュ値において，下に引数のハッシュ値の重なる部分と重ならない部分でビット'1'の立つ確率を変化させる方法を提案する．この方法によるSSCWの構成方法を第４図に示す．第４図では，関数子のハッシュ値において，下に引数のハッシュ値の重ならない部分ではαの確率でビット'1'を立て，下に引数のハッシュ値の重なる部分では，βの確率でビット'1'を立てている．実験的にαは 0.4～ 0.5，βは 0～ 0.2 程度が良い事が判明している．この方法を用いて３．(b) の測定を行なった結果を第５図に示す．この手法を用いると対象とする項の中に変数が多くなっても選択性能の大きな低下は生じなくなる．



第3図 項' a（X，Y）'のＳＳＣＷ

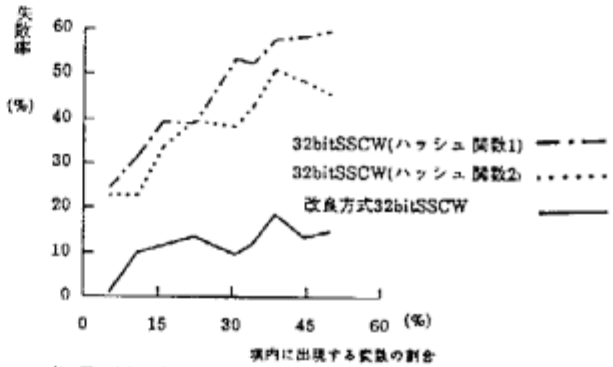


第4図　ＳＳＣＷ作成方法の改良案

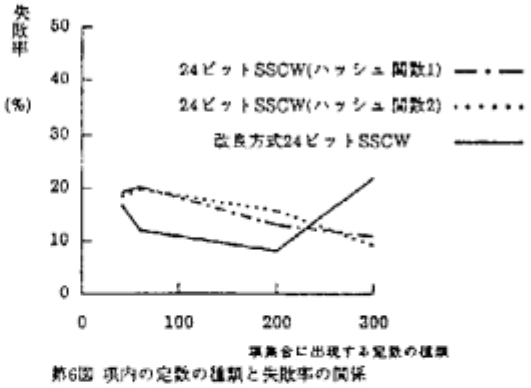
この改良案を用いると，各定数のハッシュ値の分解能力は，従来のSSCWと比較すると低下し，ハッシュの衝突も頻繁に生じるようになる．改良案によるSSCWを用いて３．(a) の測定を行った結果を第６図に示す．改良案では項集合内の定数の種類が多いと選択性能が低下するが，これは変数がSSCWに及ぼす性能低下よりも小さいことが分る．改良案において，α＝0.5，β＝0 とした場合は，[2] と等価である．α，βの最適値については，項の形状と項に出現する定数の種類，またその両者の相互関係に依存すると思われる．

6．　終わりに

以上様々な特性の項集合に対するSSCWにの選択性能の変化と，その作成方式の改良案について述べた．本稿においては，項の選択を行う時間を一定として選択性能のみについての評価を行なったが，実際には各インデックスの作成時間をも考慮して性能を評価しなければならない．インデックス作成の時間を含めた選択性能の総合評価は，今後の課題である．



第5図 項内の変数の割合と失敗率の関係



第6図 項内の定数の種類と失敗率の関係

[1] 森田他，MPPHを用いた知識ベースマシン（3）
第３５回情処全大2C-7.1987

[2] Wise.M.J.,Powers.D.M.W.,"Indexing PROLOG Clauses via Superimposed Code Words and Field Encoded Words", In Proceedings of the IEEE Conference on Logic Programming. Arrantic City. NJ. January 1984. pp.203-210.