情報検索に適した符号：順序保存符号

7H-1

中津 楢男
愛知教育大学

1. はじめに　一般に大量のデータを効率よく貯蔵するためにデータ圧縮が行われる。例えばハフマン符号はその1例であり、確かに圧縮率は最大になることが保証される。しかしハフマン符号化されたデータを検索する場合は普通はまず復号化を行い、元データを復元してから従来の検索操作が施される。この復号化時間はそれ自体微少なものであるがデータがアクセスされるたびに必要な操作であり決して無視できない。本稿では順序保存符号を用いることによって圧縮率はハフマン符号化に比べ若干劣るものの、復号プロセスのいらない高速検索が実現できることを示す。この結果は探索木の構成や2分探索されるソート済みファイルの構成に有効である。

2. 基本事項　符号は瞬時に復号可能なものだけを考え，符号語は0,1系列とする。[定義1] 情報源シンボルxの符号語をc(x)とする。情報源シンボル上に定義された全順序を≪、2進系列上の順序を<とする。任意の2つのシンボルx, yに対しx≪y ←→ c(x)<c(y)となる符号cを順序保存符号という。

[定義2] 情報源シンボルx_iの生起確率p_iが与えられた時$L=\Sigma p_i|c(x_i)|$を平均符号長と言う。ここでSは情報源、$|c(x_i)|$はx_iに与えられた符号語長（ビット）である。Lが最小となる順序保存符号を最適順序保存符号という。

最適順序保存符号を計算するためのO(n log n)時間アルゴリズム（nはシンボル数）[1]と平均符号長の理論的上限が知られている[2]。

3. 順序保存符号の応用　キー属性以外の属性値を指定して該当するレコードを高速に検索するためには2次索引を構成する。2次索引としてはB木やISAMなど探索木が使われる。探索木に必要な操作は格納されている文字列と指定された文字列の比較操作だけであり、文字列の内容は必要ない（図1参照）。この場合探索木に格納されている文字列を順序保存符号で符号化すれば、文字列間の比較を符号語間の比較で代用できる。また符号化した方が比較に必要な時間も短縮されるため、記憶量と比較時間の両方で有利である。

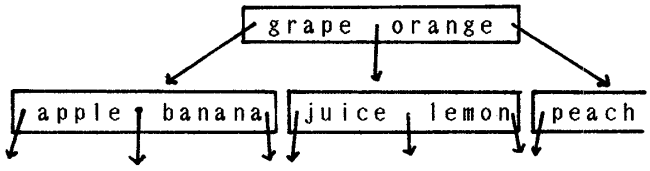

図1　探索木の例

順序保存符号の有効性を調べるために圧縮率の実験を行った。キーワード集合8696語について順序保存符号を構成したところ，その平均符号長は4.43bit/symbolであり、ハフマン符号（4.25bit/symbol）に比べ4.2%冗長であるだけであった。

[参考文献] [1]Hu,T.C.&Tucker,A.C.,SIAM J. Appl. Math.,21,4,514-532,1971
[2]中津、「順序保存符号について」、信学技研報告，IT89-56,1989.

An Order Preserved Code Suitable for Information Retrieval

Narao NAKATSU
Aichi University of Education