情報検索システムにおける文書参照ファイル

7L－1

小出 東洋† 林 雅樹‡ 竹田 正幸† 松尾 文碩†
† 九州大学大学院システム情報科学研究科 ‡ 九州大学工学部

1. まえがき

情報検索システムのファイルは，通常，文書そのものを格納した文書ファイルと，文書に対する索引部である転置ファイルとによって構成される．転置ファイルは，索引と文書参照ファイルから成る．文書参照ファイルは，索引の見出し語であるキーワードがもつ文書番号等の線形リストを格納したファイルである．文書参照ファイルでは，線形リスト長の分布に著しい偏りがある．高頻度キーワードの線形リストは格納文書数に比例して増大し，大規模システムでは数十万以上になるのに対し，約半数のキーワードのリスト長がシステム規模に無関係に1である．したがって，長短の線形リストを同一形式で2次記憶に蓄積した場合，低頻度キーワードのリストを記憶するために大量の無駄領域が生じる．本稿では，この線形リスト群の効率的管理法について論じる．

2. リスト管理法

まず，高頻度キーワードについては，線形リストの文献番号の差分をとり，それを順位符号で表現する圧縮法を開発しており，九州大学大型計算機センターにおいてINSPECテープ[1]の検索サービスを行っている情報検索システムAIR[2]で採用している．一方，低頻度キーワードについては，リスト長の同じものをまとめることによって，無駄領域を減らす方法を開発している[3]．しかし，次の点に関しては明らかにしていない．

(1) 生起頻度の上位r語を高頻度キーワードとしてその線形リストを圧縮する場合のrを定める問題．

(2) 生起頻度m以下の語を低頻度キーワードとして生起頻度ごとに管理する場合のmを定める問題．

(3) (1),(2)のいずれにもあてはまらないキーワード群に対する線形リストの管理法．

(1)については，圧縮率が問題である．圧縮率は，線形リスト中の文献番号の差分値の分布に依存するが，本稿ではこの問題に立ち入らない．

(2)については，この管理法では，文書の追加によって低頻度キーワードの生起頻度が増加したとき，線形リストを別の領域へ移動しなければならない．そのコストは，移動するリストの個数に比例すると考えられる．mをあまり大きくとると，このリストの移動が頻繁に起こることになる．リストの移動率とmとの関係は文献[3]で論じた．

(3)の問題に関しては，常識的な対処法は，物理的なアクセス単位である物理ブロックを分割した論理ブロックを用いる方法であろう．このとき，論理ブロックは大きさの異なるものを数種類用意し，低頻度なものほど小さい論理ブロックを割り当てるようにすれば無駄領域を小さくすることができる．しかし，この方式では，幾度かの文書追加により，一つの単語に対する線形リストが複数の物理ブロックに散在する事態を招き，検索時の処理速度を低下させることになる．そこで本稿では，低頻度キーワードのリスト管理法を拡張した新たな方式を提案する．

3. 中頻度キーワードのリスト管理法

非負整数の(狭義)単調増加列

$$0 = b_0 < b_1 < b_2 < b_3 < \ldots$$

を考え，リストの長さnが

$$b_{i-1} < n \leq b_i$$

On document reference file in information and retrieval system

Haruhiro Koide, Masaki Hayashi, Masayuki Takeda, and Fumihiro Matsuo
† Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka, 812-81 Japan
‡ Faculty of Engineering, Kyushu University, Fukuoka, 812-81 Japan

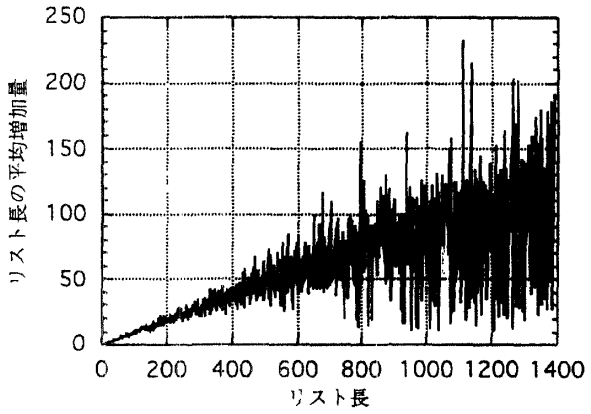

図 1 リスト長の平均増加量

となるリスト群をiごとにまとめて管理することにする．ここで，$b_i - b_{i-1} = 1$ $(i = 1, 2, \ldots, m)$ となるように数列 $\{b_k\}$ を選べば，この範囲については，低頻度キーワードに対する管理法[3]と一致する．$i > m$ なる i について，間隔 $b_i - b_{i-1}$ を大きくとれば，文書追加時に移動するリストの個数を減らすことができるが，一方，無駄領域は増加する．

延べ単語数 T の文書集合に延べ単語数 ΔT の文書集合を追加した場合に，長さ n の線形リストの長さの増加量 Δn は単語によって異なるが，その平均値 $\overline{\Delta n}$ は，n に比例すると考え，

$$\overline{\Delta n} \propto n\Delta T/T$$

とする．図 1 に，1969 年から 1984 年までの INSPEC テープ 16 年分の物理学関係の文献データに，翌 1985 年のデータを追加したときの様子を示した．ここに，$T = 187,534,540$，$\Delta T = 17,860,143$ である．

いま，長さ n の線形リストの長さの増加は平均的に n に比例すると考え，n の属する区間 (b_{i-1}, b_i) が，n に比例した“余裕”をもつようにしたい．そこで，単純に次のようにおく．

$$b_0 = 0, \qquad b_i - (b_{i-1} + 1) = \lfloor ab_{i-1} \rfloor$$

ここで，a は比例定数．

例えば，$a = 0.10$ のとき，数列 $\{b_k\}$ の先頭部分は次のようになる．

i	1 ∼ 10	11 ∼ 15	16 ∼ 19	20 ∼ 21
$b_i - b_{i-1}$	1	2	3	4

前述の 16 年分のデータに 1 年分を追加したときの移動リスト数を総リスト数 386,676 で割った移動率と無駄

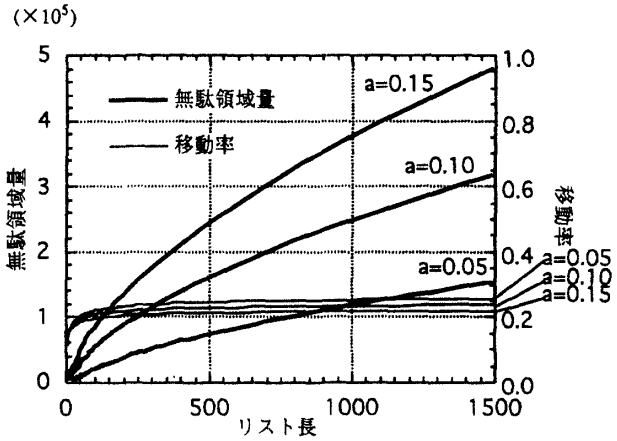

図 2 移動率と無駄領域量の累積

領域量の累積を図 2 に示した．無駄領域量の単位は，文献番号の個数で表した．図より，a の値が変化しても移動率はあまり変化しないことがわかる．これは，大多数のキーワードが低頻度語であることによる．一方，無駄領域量は，16 年分のデータにおける総リスト長 98,011,928 に比べ，十分小さい．

4. むすび

情報検索システムの文書参照ファイルにおける文書番号リストの管理法について述べた．中頻度キーワードに対する文書番号リストの管理のために，低頻度キーワードに関する効率的管理法を自然に拡張した方式を提唱した．数列 $\{b_k\}$ の決定のために定数 a の値を定める必要がある．図 2 に示した範囲では，a の値が変化しても，移動率と無駄領域量はともにあまり変化しない．一方，数列 $\{b_k\}$ の項数が多いと管理上好ましくないので，このことを考慮して a の値を決定したい．

参考文献

1) Aithison, T.M., Martin, M.D. and Smith, J.R.: Developments towards a Computer Based Information Services in Phisics, Electrotechnology and Control, *Inform. Strage and Retrieval*, 4(2), 177–186 (1968).

2) Matsuo F., Futamura, S. and Shinohara, T.: Efficient strage and retrieval of very large document databases, *Proc. of the 2nd Int. Conf. on Data Eng.*, 456–463 (1986).

3) 松尾，佐藤，高山: 情報検索システムにおける文書参照ファイルの効率的構成 **36**(6), 1486–1494 (1995).