2H－2

ハッシュ技法によるデータ検索の数学的モデル化とその評価 -キーワードによる文献検索モデルに対して-

二神常爾，松嶋敏泰，平澤茂一
早稲田大学理工学部経営システム工学科

1　まえがき

キーワードを用いた文献検索では個々の文献をキーワードの集合によって特性づける．簡単なモデルではフォーマット化を行い個々の文献 j に長さ N の二次情報ベクトル $\vec{x}_j$ を対応させる．個々のビット位置はキーワードの有無に対応し，1のあるビット位置のキーワードは文献に含まれるが，0のビット位置のキーワードは文献に含まれないことを意味する．質問(query)$\vec{x}$としていくつかのキーワードを与えて，これらのキーワードをすべて含む文献の二次情報レコード $\vec{x}_j$ を求める．これを $\vec{x} \subseteq \vec{x}_j$ により表す．次にこの二次情報レコードに対応する一次情報を二次記憶内から探し出すのが通常の文献検索である．

上の簡単なモデルでは，二次情報の大きさは文献数 n_0 と二次情報レコードの長さNの積になるがこれが主記憶（RAM）容量よりもかなり大きい場合には二次情報を縮小してRAMに格納せねばならない．これはRAMに格納する1レコードあたりの情報がキーワードの総数 N よりも小さいことを意味する．この場合に $\vec{x} \subseteq \vec{x}_j$ を満足する $\vec{x}_j$ をすべて求めようとすると，必然的に不必要な $\vec{x}_j$（$\vec{x} \subseteq \vec{x}_j$ を満たさない）の一部分も含まれてしまう．この結果，必要な文献の一次情報ばかりでなく不必要な文献の一次情報も二次記憶装置からRAMにロードすることになる．これは二次記憶装置へのアクセス回数が増大することを意味している．したがって，不必要な $\vec{x}_j$ の総数（extratransmission，余剰転送数と呼ぶ）をできるだけ低減することが重要である．余剰転送数を低減するためにRAMの情報を利用する考えは参考文献[1]で報告されている．参考文献[2]ではベクトル $\vec{x}$ とのハミング距離がしきい値 T 以内にあるレコード $\vec{x}_j$ を求めることを目的とする場合に，余剰転送数の上界の評価が情報理論の視点から行われた．

本研究の一部は文部省科学研究費 基礎研究B07558168，早稲田大学特定課題研究の助成による．

"Document retrieval models using keywords."
T.Futagami,T.Matsushima,and S.Hirasawa
Department of Industrial and Management Systems Engineering, School of Science and Engineering, Waseda University
e-mail: futagami@hirasa.mgmt.waseda.ac.jp

2　提案モデル

ここでは，RAMの情報を $\vec{x}_j$ 1個あたり N よりも小さくする方法として以下の方法を考える．長さ N の文献レコード $\vec{x}_j$ を l ビットごとに区切る（$l|N$）．この連続する l ビットをブロックと呼ぶ．個々のブロック内に1が一つ以上あれば1を対応させ，すべて0であれば0を対応させる．こうしてつくられる N/l 個の0,1の系列を $\vec{V}(\vec{x}_j)$ として，RAMに $\vec{x}_j$ の代わりに $\vec{V}(\vec{x}_j)$ を記憶させる．RAM内の情報の大きさは元の二次情報の大きさの $1/l$ 倍になっている．このとき $\vec{V}(\vec{x}) \subseteq \vec{V}(\vec{x}_j)$ は $\vec{x} \subseteq \vec{x}_j$ の必要条件となっている．RAMの二次情報を用いて $\vec{V}(\vec{x}) \subseteq \vec{V}(\vec{x}_j)$ を満足する $\vec{x}_j$ の一次情報を二次記憶装置からロードするときには必要な $\vec{x}_j$（$\vec{x} \subseteq \vec{x}_j$ を満たす）ばかりでなく不必要な $\vec{x}_j$（$\vec{x} \subseteq \vec{x}_j$ を満たさない）の一次情報の一部もロードされる．ちなみに，一次情報の格納される二次記憶内のハッシュアドレスは

$\vec{V}(\vec{x}_j)$ を用いて決める．ハッシュ空間の大きさは $2^{N/l}$ であり，これは $\vec{x}_j$ の総数 n_0 にほぼ等しくなるように選ぶ．

3 解析，評価

さて，上のモデルを用いて RAM のメモリ量 R_M と余剰転送数 n_e（$\vec{V}(\vec{x}) \subseteq \vec{V}(\vec{x}_j)$ は満足するが $\vec{x} \subseteq \vec{x}_j$ は満足しない $\vec{x}_j$ の個数）の関係を求める．N/l 個の各ブロックにおける 1 の個数を $(r_1, r_2, \ldots, r_{N/l})$ によって表す．また，各ブロックでの 1 の個数が $(u_1, u_2, \ldots, u_{N/l})$ となる $\vec{x}_j$ の個数を $n(u_1, u_2, \ldots, u_{N/l})$ とする．ここで，次式が成り立つ．

$$\sum_{r_1=0}^{l} \cdots \sum_{r_{N/l}=0}^{l} P(r_1, r_2, \ldots, r_{N/l}) = 1 \quad (1)$$

$$\sum_{u_1=0}^{l} \cdots \sum_{u_{N/l}=0}^{l} n(u_1, u_2, \ldots, u_{N/l}) = n_0 \quad (2)$$

余剰転送数の平均値を $\bar{n}_e$，必要な $\vec{x}_j$ の数の平均値を $\bar{n}_{true}$ により表すと

$$\bar{n}_e + \bar{n}_{true} = \sum_{u_1=r_1}^{l} \cdots \sum_{u_{N/l}=r_{N/l}}^{l} P(r_1, \ldots, r_{N/l}) \cdot \sum_{u_1=r_1}^{l} \cdots \sum_{u_{N/l}=r_{N/l}}^{l} n(u_1, \ldots, u_{N/l}) \quad (3)$$

ここで以下の条件下での $\bar{n}_e$ を求める．まず，次のように $P(r_1, \ldots, r_{N/l})$ を仮定する．

$P(r_1, \ldots, r_{N/l}) = 1/{}_{N/l}C_s$
（s 個の r_i に対し $r_i = 1$，
他の r_i に対し $r_i = 0$ の条件下）

$P(r_1, \ldots, r_{N/l}) = 0$，（他の条件下）

また，$\vec{x}_j$ の N 個のビット位置での 1 の生起確率は等しく p であるとする（$p \ll 1$）．すると $\vec{x}_j$ の分布として次式を得る．

$$n(u_1, \ldots, u_{N/l}) = n_0 {}_lC_{u_1} \cdots {}_lC_{u_{N/l}} \cdot p^{u_1+\cdots+u_{N/l}} \{1-p\}^{N-u_1-\cdots-u_{N/l}} \quad (4)$$

これらの仮定に対し次式を得る．

$$\bar{n}_e + \bar{n}_{true} = n_0 \{1 - \{1-p\}^l\}^s \quad (5)$$

$$\bar{n}_{true} = n_0 p^s \quad (6)$$

ちなみに $l = 1$ の場合は RAM に格納する情報を縮小しない場合であるが，このときは当然ながら $\bar{n}_e = 0$ となる．$pl \ll 1$ の場合には次式が成り立つ．

$$\bar{n}_e = n_0 p^s (l^s - 1) \quad (7)$$

さらに次式が成り立つ．

$$R_M = N/l \quad (8)$$

式 (7),(8) より次式を得る．

$$\bar{n}_e = n_0 p^s \{(N/R_M)^s - 1\} \quad (9)$$

4 まとめ

余剰転送数を低減するために RAM の情報を利用する考え方を文献検索に対して適用できることを示し，簡単なモデルに対して R_M と n_e の関係を導くことができた．$\vec{x}$ と $\vec{x}_j$ のより一般的な分布に対して R_M と n_e の関係を求めるのが今後の課題である．

参考文献

[1] G.H.Gonnet and P.A.Larson, "External hashing with limited internal storage.", J.ACM, vol.35, no.1, pp.161-184,1988.

[2] V.B.Balàkirsky, "Hashing of Databases based on indirect obserbations of Hamming distances.", IEEE Trans. Inform. Theory, vol.42, no.2, pp.664-671, 1996.