高能率2次文献情報検索システムの設計

5E－10

高山 悟† 高松 希匠†† 佐藤 誉夫††† 竹田 正幸† 松尾 文碩†
†九州大学工学部 ††シャープ(株) †††日立超LSIエンジニアリング(株)

1. まえがき

現在，九州大学計算機センターでは，AIRとよばれる情報検索システムによってINSPEC検索サービスが行われている．AIRは富士通のIBM互換OSのもとで動作し，大部分はFortranで書かれている．しかし，Fortranのコンパイラの改版に伴い，再コンパイルができなくなっており，将来のサービス継続が危ぶまれている．そこで，著者らが開発したSeepによってAIRを再構築することにした．SeepはC言語によって実現されていて，UNIXあるいは，MS-DOSのもとで動作する．Seepは，ルールベースとデータベースを統合的に管理するシステムであるが，基準的データベース管理技法によって実現しているため，一般的なDBMS同様，数十万以上の文献を効率的に管理することができない．ここではSeepにAIR並の数百万以上の文献を管理する能力をもたせるための方策について述べる．

2. Seep

Seepは，ルールベースとデータベースを統合管理するシステムであり，特徴としてインタプリタを持たず，ルール適用の制御をコンテクスト（作業メモリ）によって行うという方式を採用している．Seepのデータベースの管理機構の外部インタフェースはDMP (*data manipulation primitives*) である．DMPはFIND，INSERT等のデータベースの基本操作の集合であり，C言語で記述された副手続きである．これらの副手続きを応用プログラムあるいはルールから呼び出すことにより，データベースを操作することができる．

A design of an efficiency information retrieval system
Satoru Takayama†, Kisyo Takamatsu††, Takao Sato†††, Masayuki Takeda† and Fumihiro Matsuo†
†Kyushu University 36, Hakozaki, Fukuoka, 812 Japan.
††SHARP Corp. †††Hitachi ULSI Engineering Corp.

3. 不要語除去

文章の大半を占め文書内容の識別力がほとんどない機能語を，ここでは不要語とよぶ．また不要語の辞書をあらかじめ持っておいて，その辞書にない語を索引語とする自動索引を不要語除去法とよぶ．現在，大量書誌的文献に対する自動索引には，不要語除去法以外に実用に耐える方式がない．ここでは，対象となるINSPECテープの書誌項目である自由索引句 (*free-indexing terms*) と抄録 (*abstract*) に生起する単語 w の生起確率 $f_i(w)$，$f_a(w)$ の比 $f_i(w)/f_a(w)$ を不要語選択の基準として利用する．この方法によって，問題とされる検索能力と転置ファイルの大きさを考慮に入れた自動索引が可能になる[3] この方法に従って，25年分のINSPECテープについてデータを取り直した．

4. 転置ファイルの構成

情報検索システムにおいて参照されるファイルは，通常，文書そのものを格納したファイル (*document file*) と，索引部となる転置ファイル (*inverted file*) とから構成される．ここでは，このうち情報検索システムの性能に最も関係しているキーワードに関する転置ファイルを対象にする．

転置ファイルは図1のように転置索引 (*index*) と文書参照ファイル (*document reference file*) とよばれる2つのファイルから構成され，ここで用いるキーワード転置ファイルは，単純転置ファイルとよばれるもので，キーワードの内容がそのキーワードを含む文書番号の線形リストになっている．その，文書参照ファイルとよばれる線形リストのファイルは英単語の生起頻度の特徴を反映して高頻度キーワードと低頻度キーワードの場合とで，そのリストの長さの分布に著しい偏りがある．このため2次記憶領域の効率的確保という観点からは，長短の極端なこれらのリストを混在させて管理することはできない．このため，それぞれの場合に分けて，効率的な2次領域確保のための文書参照ファイルの構成法が必要になる．

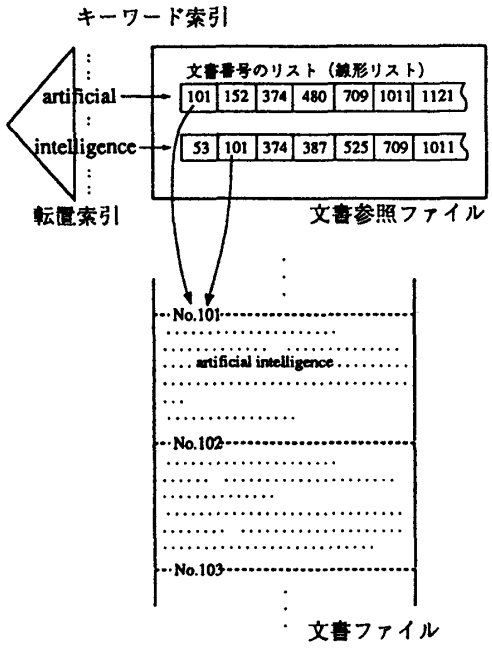

図 1 転置ファイルの構成

5. 高頻度キーワード

転置索引の見出し語（キーワード）が高頻度の場合，その線形リストには数万から数十万以上もの文書番号が存在する．これらの文書番号を，例えば 4 バイトなどの固定長符号で表すのはかなり冗長であり，2 次記憶領域の圧迫要因となるため，これの圧縮が重要である．まず，文書番号の代わりに前の番号との差分をおいて数の桁数を小さくし，更に，この差分値リストを順位符号という可変長符号によって圧縮する[4) 5)]．

6. 低頻度キーワード

ここでは，データ追加による追加領域が問題となる．低頻度キーワードにおいては，英単語の生起確率について有名である Zipf の法則が当てはまらず，次式が成立することが示されている [2)]．

$$p(r) = c/r^2, \qquad (1)$$

ここで，c は定数.

文書追加の際，個々の低頻度単語 w の増加を予測することはできず，そのため，追加領域を効果的に確保できない，しかし (1) 式から，既知であるところの文書量，追加文書量を使って，生起回数毎に単語を集合化すると，その集合の増加量は予測できることがわかる[1)2)]．

そのことを利用して，生起毎の線形リスト群に分割する方式が考えられる．この方法では，無駄な追加領域が生じない．

つまり 2 次記憶領域の効率的利用という点で大きく改善される．しかし，この方式では文書の追加において，あるキーワードの生起回数が増加したとき，そのキーワードの線形リストを別のリスト群に移動しなければならない．その移動コストは，移動するリストの数に比例すると考えられる．

この場合，どの生起回数の語までまとめるかという m は理論的に決定できない．m の増加に伴って，領域量は減少し，計算量は増加するが，計算量の増加はデータ追加にかかわることであり，検索性能とは無関係であるため，領域量の減少と同列に比較する問題ではないからである．したがって，m を決める問題は，情報検索システムの実現と管理運用の容易さと領域量の減少の程度との兼合いで決まることになる[6)]．

7. まとめ

本稿で述べた文書参照ファイル構成法を Seep に組み込むことにより，Seep は高能率 2 次文献情報の検索に関し AIR 以上の性能が得られるものと考えられる．現在，Seep を使った新 AIR の構築を行っている．

参考文献

1) Booth, A. D. : A “Law” of Occurrences for Words of Low Frequency, Inform. Control, Vol.10, No.4, pp.386-393 (1967).

2) Matsuo, F. : On Word Occurrence in Scientific and Technological Texts, 情報処理学会自然言語処理研究会資料 46-2 (1984).

3) 二村祥一，松尾文碩：英文科学技術文献情報に対する不要語除去法による自動索引，情報処理学会論文誌，第 28 巻，第 7 号，pp.737-747 (1987).

4) 二村祥一，松尾文碩：順位符号に基づく英文二次文献情報のデータ圧縮法，情報処理学会論文誌，第 28 巻，第 3 号 (1987).

5) 佐藤誉夫，松尾文碩：情報検索システムにおける文書番号リストの圧縮，第 46 回電気関係学会九州支部連合大会講演論文集 (1993).

6) 松尾文碩，佐藤誉夫，高山 悟：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，情報処理学会論文誌 第 36 巻，第 6 号 (1995).