7H－7　キーワードの登録回数に基づくグループ化による一検索方式

菊池忠一(1)　藤村浩(2)　藍沢實(1)
(1)(株)テレマティーク国際研究所　(2)電気通信大学

1　はじめに

電子メディアの普及と通信の高度化に伴い、電子図書館の構築が進められるようになってきた。図書には、著者／発行者／件名等の数種のキーワードがあり、検索時にはこれらの中から任意のキーワードを使用されるのが普通である。また、一般的な数万冊から約１００万冊蔵書を有する図書館の場合、現状、検索に時間がかかり、検索を煩わしいものにしている。

現在図書検索に用いられている検索手法としては次のものがある。①逐次サーチ法は検索ファイルのデータ量に比例した検索時間を要し、大量情報向きではない。②転置ファイル法は、検索入力数に比例した検索時間を要し、複数検索入力向きではない。これらの改良方式として、特性ファイル法のフォルスドロップを改良した重ね合わせ符号法がある。これは、システムの保守・管理が転置ファイルより容易で、逐次サーチ法より高速検索できる[1]が、大量情報検索には必ずしも十分とは言えない。

本報告では、大量情報における高速検索方式、特に、複数の検索入力における論理積条件下での高速検索方式を提案している。４万冊の書誌情報を対象とする検索実験で、平均プログラム実行時間２．３msの検索速度を確認した。

2　アルゴリズムの特徴

本方式は、複数のキーワードを使用し、大量情報の中から高速検索できるもので、以下の特徴を持つ。

(1)　登録使用頻度の少ないキーワードから検索する方が高速性が得られる特徴を活かした検索ファイル構造である。
　①　検索ファイルには、キーワード登録使用頻度の低頻度から昇順に、全キーワードに対応した検索レコードがわりあてられる。
　②　各登録情報の有するキーワードについて、検索レコードに対応するキーワードの頻度より高頻度のキーワードをすべて格納する。（図１）

(2)　検索ファイルは(1)の①により、各検索レコード容量が均等化する方向に作成される。

(3)　複数のキーワードによる検索の場合、低頻度のキーワードに対応する検索レコードだけで、検索可能である。

(4)　キーワード種ごとにマーク付けし、すべて４バイトでコード化することにより、キーワード種別処理が不要である。

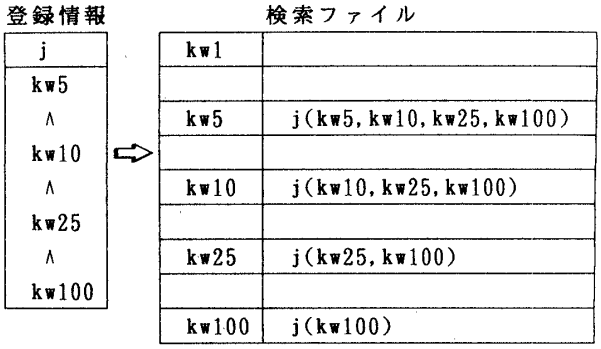

図１　キーワード登録例

3　登録

3．1　キーワード表の作成

著者名と発行者名にそれぞれ専用マークを付加し、件名も含めた３種のキーワードを登録順に、１から昇順の数値を付与してコード化する。コード化と同時に登録回数を計数し、キーワード表（図４）を作成する。例えば図４では、ＩＳＤＮはキーワードとして５番目に登録され、既にＩＳＤＮをキーワードに持つ書誌情報が１００件あったことを示す。

3．2　検索ファイルの作成

検索ファイル（図２）に、キーワード登録回数の最も少ないものから昇順に、すべてのキーワードに対応した検索レコード（図３）を割り当て、登録情報の有するキーワードの中から、検索レコードに対応するキーワードの登録回数より多いキーワードをすべて格納する。例えば、ＩＳＤＮ，通信，ＯＳＩを有する書誌番号１０の場合、キーワード表から登録回数順はＯＳＩ＜ＩＳＤＮ＜通信であるから、ＯＳＩのコード１５０に対応する検索レコードにＩＳＤＮ，通信，ＯＳＩを格納し（図３）、ＩＳＤＮの

A Group Retrieval Technique Based on Registration Frequency of Keyword

Chuichi Kikuchi(1)　Hiroshi Hujimura(2)　Minoru Aizawa(1)
(1)Telematique Intl. Res. Lab.　(2)University of Electro-Communications

コード５に対応する検索レコードにはＩＳＤＮと通を格納し、通信のコード２０に対応する検索コードには通信を格納する。

4　検索

キーワード表を使用して、入力されたキーワードの内、登録回数の最も少ないコードをレコード番号とする検索レコードから、入力されたすべてのキーワード（コード）が共通に持つ書誌番号を取り出す。

例えば、図３と図４において、検索入力がＩＳＤＮ，通信，ＯＳＩの場合、登録回数が最も少ないＯＳＩに該当するコード１５０をレコード番号として、検索レコードを取り出す。この検索レコードの中から、ＩＳＤＮ，通信，ＯＳＩのコード５，２０，１５０に共通する書誌番号として１０を取り出す。

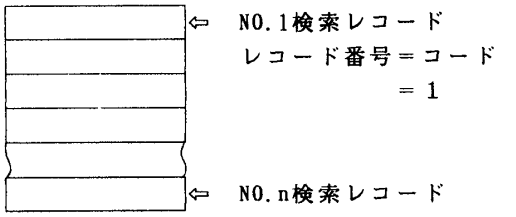

図２　検索ファイル

レコード番号	150
コード 5	書誌番号 10
5	78
20	3
20	10
20	222
150	10

図３　検索レコード

キーワード	コード	登録回数
ＩＳＤＮ	5	１００
ソフト	１７	３５
通信	２０	３５０
ＯＳＩ	１５０	５０

図４　キーワード表

5　性能試験

表記方式を評価するために、ＰＦＵ社製Ａ－５０ミニコン（ＯＳ：ＵＮＩＸ）上に、４００３６冊の書誌情報とキーワードを登録した実験システムを構築し、登録時間，検索ファイル容量，書誌情報が有する全キーワード検索の時間を調べた。

使用した書誌情報（表１参照）は、キーワード（著者名，発行者名，件名）が１個の書誌情報が６４件，２個の書誌情報が１２１８件，・・・，１２個の書誌情報が３３件である。キーワードは合計４８００８種あり、１書誌情報平均４．８個である。

試験結果は、検索ファイルは約４．４ＭＢ，キーワード表は約１．７ＭＢ，検索ファイル・ポインタ表は約０．５ＭＢであった。登録時間は約５．６時間，検索時間は表１に示すとおりであった。なお、検索時間は、検索ファイルから該当する全書誌番号を取り出しまでの時間で、ディスク装置アクセス時間を除くプログラム処理時間である。

検索入力	検索時間			検索回数
キーワード数	平均 ms	最小 ms	最大 ms	書誌情報数
1	1．5	1．1	8．8	64
2	1．6	1．3	5．9	1218
3	1．8	1．4	6．8	6324
4	2．0	1．6	11．8	12401
5	2．3	1．8	12．9	9422
6	2．7	1．9	12．9	5109
7	3．0	2．1	11．4	2865
8	3．3	2．3	10．2	1500
9	3．4	2．5	7．7	656
10	3．3	2．7	5．7	304
11	3．4	2．9	5．1	140
12	3．5	3．0	4．9	33
全体	2．3	1．1	12．9	40036

表１　検索時間

6　おわりに

約４万冊書誌情報を対象とする性能試験で、平均検索時間２．３msを確認できた。これはプログラム実行時間のみで、実際にはこれにディスク装置アクセス時間（約３０ms）を加えたものになる。

表１の検索結果から平均検索時間と最小検索時間は、検索入力されたキーワード数ｎの一次関数で表わせることがわかる。

平均検索時間　$t = 0.26(n-1)+1.3$　ms　$n \leqq 8$

最小検索時間　$t = 0.18(n-1)+1.08$　ms　$n \leqq 12$

係数は、該当検索レコード内で検索入力キーワードに共通する書誌番号を抽出する論理積にかかる時間である。

平均検索時間に着目すると、$n \geqq 8$まではほぼ比例するが、$n \geqq 9$では飽和してくる。これは、キーワード８個前後で検索結果が出ていることと、書誌情報が急減していることによると考えられる。

本方式は、複数の検索入力に対し、ディスク装置から１個の検索レコードを取り出すだけで検索可能なことから、複数のキーワードを持つ大量情報の高速検索に有効な方式であると考えられる。

今回は、約４万冊を対象に性能試験を行い検索の高速性を確認したが、今後、本方式の特徴を明確にするために大量情報の検索を試みたい。

参考文献

[1] 有川，篠原他：「重ね合わせ符号を用いた文献検索システムについて」：情処研報データベースシステム５４－２