2N-5

キーワードのコード化による一検索方式

菊池忠一　飯島豊
(株)テレマティーク国際研究所

1．はじめに

最近、電子メディアの普及が著しいが、情報検索時に入力するキーワード（以下ＫＷと略す）がわからないとか、ＫＷの一部しかわからないという場合が多く、検索を煩わしいものにしている。

検索時に書名や見出し語の一部、時には単語の一部しか記憶していない場合があり、この場合、書名や見出し語の一部でも検索可能な部分一致や中間一致検索が有効になる（部分一致＝前方一致∪中間一致∪後方一致）。しかし、現行の方式では転置ファイルが大きくなるために、部分一致や中間一致検索は利用できないことが多い。この解決の一手法として、インデックスを用いない国語辞書全文検索方法［1］が提案されているが、この手法は検索ファイル内の全キーワードを照合する必要がある。本報告では、ＫＷをコード化して取り扱い、該当文字ブロックを照合することにより、上記５種の検索を共通の処理で実現し、あわせて検索ファイルの小容量化による検索ファイルの主記憶常駐および検索の高速化も実現できる検索方式を述べる。

2．検索方式

2．1　キーワードのコード化

ＫＷを分野別など、情報を整理利用するのに適した登録番号を付加する。このＫＷを一文字ごとに分解し、各文字に先頭からの文字位置を示す文字番号を与える。この登録番号と文字番号を式１に代入し、文字コードを作成することでキーワードをコード化する。例えば、「ＩＳＤＮ」に登録番号１２５を与える場合、文字コードは図１のようになる。

文字コード＝登録番号×１００＋文字番号・・式１

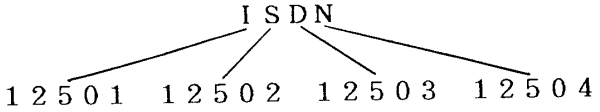

図１　ＫＷコード化の例

A Information Retrieval Methode
Using Keyword Coding-Technique
Chuichi Kikuchi , Yutaka Iijima
Telematique International Research Laboratories

文字コードを４バイトで構成すると、最大１００文字長のＫＷを、2^{32}／１００＝約４千万種、整数として取り扱うことができる。また、ＫＷ長が最大２０文字の場合には、2^{32}／２０＝約２億種のＫＷを扱うことができる。

2．2　検索ファイル

検索ファイルは、システムで使用可能な各文字に対応するブロックで構成される。例えば、ＪＩＳ第１及び第２水準漢字表（以下ＪＩＳ表と略す）の各文字をＵＮＩＸコードで取り扱う場合は、式２により変換したコードをブロック名としている。ＫＷを登録する場合には、ＫＷの各文字に対応するブロック内の末尾に、該当する文字コードを格納する。「ＩＳＤＮ」を登録する例を図２に示す。このとき、各ブロック内の文字コードは、昇順に格納されることになる。

ブロック名＝ＵＮＩＸコード－Ａ１Ａ０・・・式２

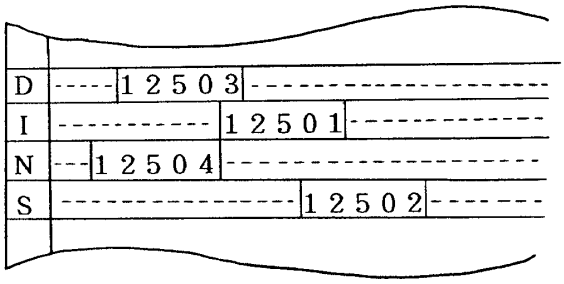

図２　検索ファイル登録例

文字コードを４バイトとすると、ファイル容量は４バイト×$\sum_{i=1}^{n}$(ＫＷ文字数)$_i$で小規模になるので主記憶常駐が可能になり、検索の高速化も実現できる。

2．3　検索方法

検索時は、入力された文字列を一文字毎に分解し、式２からブロック名を算出し、文字順に該当するブロックを検索ファイルから読み出す。次いで式３に示すように、ブロック間で登録番号が等しく文字番号が１差の文字コードを取り出す。

（文字順ｎのブロック内文字コード）－
（文字順ｎ＋１のブロック内文字コード）
＝－１・・式３

この後、検索指示に従い、前方一致／後方一致／完全一致／部分一致／中間一致特有の処理を行い、文字コードから検索対象の登録番号を取り出す。

例えば、検索入力ＩＳで前方一致検索を行う場合、図２からＩのブロックとＳのブロックを取り出し、式３を使用して該当する文字コードを抽出すると、この中に文字コード１２５０２も含まれる。次に、前方一致特有処理として、検索入力ＩＳの文字数２に等しい文字番号を有する文字コードを捜すと、この中に１２５０２も含まれ、「ＩＳＤＮ」が検索されていることがわかる。残り４種の検索も同様に文字番号の処理を変更することで実現できる。

２．４　その他の処理

追加登録は、２．１と２．２で説明した登録処理同様に、ＫＷの各文字に該当するブロックの末尾に新規文字コードを追加することで行う。また、削除は、削除ＫＷの各文字に該当するブロックの該当文字コードを特殊記号に変更し、登録番号を欠番にする。この結果、追加登録と削除を短時間で行えるようになった。

３．実験結果

標記方式を評価するために、ＰＦＵ社製Ａ－５０ミニコン（ＯＳ：ＵＮＩＸ）上に、５１０００冊の図書名を登録した実験システムを構築した。洋書の検索も考慮し、ＫＷ長を最大１００文字として式１で文字コードを作成し、実験では、検索ファイル容量、検索時間、追加登録時間、削除時間を調べた。ファイル容量は約１．８Ｍバイトで、各処理時間は表１にとおりである。なお、使用した図書名は１文字～２２文字で、平均は９文字である。また、検索時間は検索ファイルから該当する全登録番号を取り出すまでの処理時間、追加登録時間はディスクと主記憶の検索ファイルにＫＷを登録するまでの処理時間、削除時間はディスクと主記憶の検索ファイル内の登録データを特殊記号に変更するまでの時間である。

	最小	最大	平均
完全一致検索	４ms	１２１ms	７０ms
前方一致検索	２ms	１７５ms	３２ms
後方一致検索	３ms	３４５ms	３２ms
部分一致検索	３ms	３２２ms	３１ms
中間一致検索	４ms	７８０ms	３２ms
追加登録	３８Ｓ	３９Ｓ	３８Ｓ
削除	３５Ｓ	４８Ｓ	３６ｓ

表１　処理時間

４　おわりに

ＫＷをコード化して取り扱うことにより、完全一致／前方一致／後方一致／部分一致／中間一致の各種検索を共通の検索ファイルで実現できること、検索時間が完全一致を除くと平均３２ｍｓであること、また、検索ファイルがパソコンやワークステーション主記憶に展開できるほど小規模に構成できることを確認した。さらに、追加登録時間が平均３８ｓ、削除時間が平均３６ｓであり、データベース・メンテナンス面も実用的な速度と判断される。

また、本方式は、ＫＷをその登録番号と文字番号で管理できるが、これはＫＷだけではなく、情報の総称に登録番号を付加し、情報の構成部分に出現番号を与えることにより、情報一般の登録検索に応用できると考えられる。

さらに本方式は複数のブロックから構成される検索ファイルに、コード化された文字コードを格納する方式なのでハード化が容易であることから、図書や辞書検索以外の各種情報検索にも応用可能と思われる。なお、意図的に平仮名表現がされている、あるいは、複数の読みが可能なＫＷを対象とする検索への適用が今後の課題である。

５　謝辞

本テーマの取り組みにあたって助言をいただいた当研究所第一研究部卯月前部長（現東京電力システム研究所主席研究員）、日頃から御指導いただいている第一研究部藍沢部長および有益な議論をいただいた山田主任研究員はじめ第一研究部各位に感謝いたします。

参考文献

［１］　福島・菊地、文字列検索ＬＳＩを用いた国語辞書システムの構築法、情処３７全大５Ｂ－１０、１９８８