自由語検索のための高速文字列検索方式

2N-8

川口　久光　　加藤　寛次　　藤澤　浩道　　畠山　敦　　藤縄　雅章

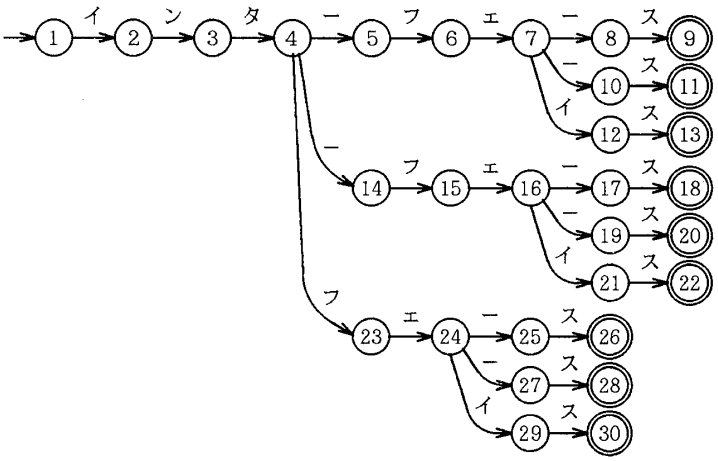
（株）日立製作所　　中央研究所

1．はじめに

文書情報検索システムでは，文書の検索に際して，ユーザが目的の文書について思い付いた自由な言葉（自由語）で検索できる検索システムが求められている。このためには，同じ意味を表わす同義語および表記の異なる異表記語を考慮した検索を行う必要がある。

これを実現する方式として，ユーザが投入した検索語から同義語・異表記語を自動的に生成し，これら数百の検索語を一括して照合する自由語検索方式を提案している。［１］　この自由語検索において，必要とされる複数検索語の一括検索方法として有限オートマトンを用いたハードウェアを開発してきた。［２］

本稿では，高度な検索機能を実現しようとするとオートマトンの状態数が多くなるという従来方式の欠点を改善したコンカレントステート型オートマトン方式を提案するとともに，そのハードウェア方式について言及する。

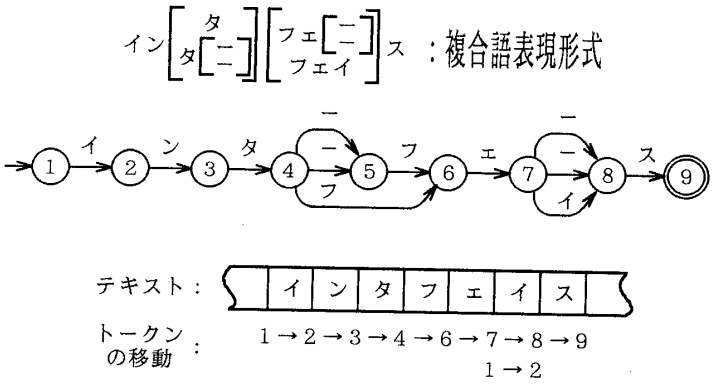


第1図　異表記検索用オートマトン（従来方式）

2．コンカレントステート型オートマトン方式

従来は，活性化しているオートマトンの状態を示すトークンを１つだけ照合処理に使用していた。このため，状態遷移のフェイル処理を考慮してオートマトンを作成しなければならないという制約があった。この制約のため木状に広がるオートマトンしか記述出来ず，異表記語を照合するオートマトンを作成しようとすると状態数が非常に多くなってしまうという問題があった（第１図参照）。

この課題を解決するために，複数のトークンを照合処理に使用することによりフェイル処理を不要としたコンカレントステート型オートマトン（ＣＳＡ）方式を提案する。本方式ではフェイル処理が不要なため，オートマトンの状態遷移を網状にまとめて記述することができる。その結果，異表記用のオートマトンも極めて少ない状態数で実現することが可能となった（第２図参照）。



第2図　異表記検索用オートマトン（ＣＳＡ方式）

A Fast String Searching Algorithm for Free Term Search

Hisamitsu KAWAGUCHI　Kanji KATO　Hiromichi FUJISAWA　Atsushi HATAKEYAMA　Masaaki FUJINAWA

HITACHI,Ltd.

３．ＣＳＡ型サーチエンジンの構成

ＣＳＡ方式に基づく検索ハードウェア（サーチエンジン）は，オートマトンを格納する状態遷移テーブル，トークンを格納する２面バッファ，および照合結果識別子を格納する照合結果テーブルから構成する（第３図参照）。第３図では、４０９６状態のオートマトンを格納することが可能である。

１回の状態遷移は，テキストから１文字とバッファＡからトークンが存在する現在の状態番号を読込み，状態遷移テーブルを参照してからトークンが移動すべき次の状態番号を読み出しバッファＢに格納することにより実現している。バッファＡとバッファＢを切り替えながら，次々と文字コードを読込み同様の処理を繰り返すことによって文字列の照合を行う。

４．高度な検索機能

ＣＳＡ方式では，従来困難であった下記①～④の高機能な検索を少ない状態数のオートマトンで実現できる。

①異表記検索：

インタフェース → インターフェース
インターフェイス
・
・
・
・

②don't care文字指定検索：

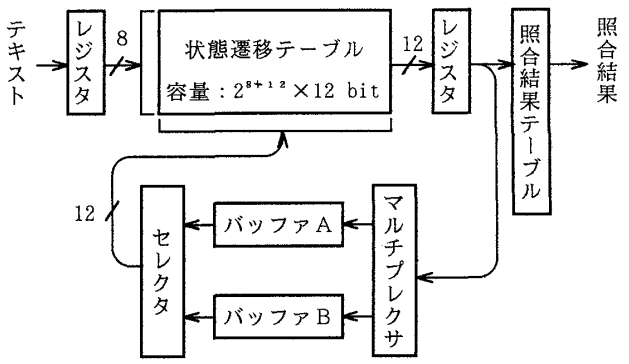
漢字？？方式 → 漢字認識方式

③近傍条件指定検索：

文字［３Ｃ］方式 → 文字認識方式
文字読取り方式

④１文字誤り許容検索：

読取り方式 → 読み取り方式
読取方式



第3図　ＣＳＡ方式サーチエンジンの構成

５．おわりに

本方式によれば，フェイル処理が不要なためオートマトンの状態遷移を網状にまとめて記述でき，従来に比べ少ない状態数で検索オートマトンを作成することが可能となった。このことにより，異表記検索，don't care文字指定検索，近傍条件指定検索，１文字誤り許容検索などの高度な検索機能を有する高速サーチエンジンを実現できるようになった。

参考文献

［１］畠山他：“知的ファイリングシステムの開発（その２）；自由語検索における異表記，異表現解消法”，第３３回情処全大４Ｙ－９（１９８６）．

［２］川口他：“高速検索文字列検索方式”，昭６２信学全大５４（１９８７）．