6Q−7　文書検索向き数値検索方式の提案

志村隆則　川口久光　加藤寛次　畠山　敦　秋沢　充
（株）日立製作所　中央研究所

1．はじめに

文書や書類を検索する場合，その文書中の数値データに関して大小比較をしたり，範囲を指定して検索したいという要求がある。一般に，文章内の数値データは左詰めになっているので，数値列の１文字目の数値を読み出した段階では，その数値列が何桁の数値か分からない。最終桁の次の文字，すなわち数値列に続く非数値文字を検出して初めて，その数値文字列が何桁かが分かる。したがって，高速な数値検索機能を実現するためには，文字列を１文字ずつ読み出しながら逐次大小比較処理を行い，数値文字列を読み終わった段階で，直ちに比較判定結果を出力できる数値検索方式が必要となる。

2．数値条件分割方式の提案

数値検索機能を実現する方法として，オートマトンを用いた方式がある。この方式では，検索したい範囲（以下，数値条件という）の数値全てに対して状態遷移パスを割り付ける必要がある。そのため，状態遷移テーブルが大規模になり，オートマトン生成時間が長くなるという問題がある。

本報告では，数値条件を桁数ごとに分割して，各桁ごとに数値検索を行なう数値条件分割方式を提案する。本方式によれば，オートマトンが簡単になるので，状態遷移テーブルが小さくなり，オートマトンの生成時間も短縮することが可能となる。

数値条件の分割方法は，以下の通りである。

（１）数値条件の下限値と上限値の桁数が２桁以上異なる場合

数値の範囲を次の３つの範囲に分割する。

①下限値から下限値と同じ桁数の最大数値まで。

②下限値より１桁大きい数値の最小値から上限値より１桁小さい数値の最大値まで。

③上限値と同じ桁数の数値の最小値から上限値まで。

（２）数値条件の下限値と上限値の桁数が１桁異なる場合

（１）の①と③の２つの数値範囲に分割する。

（３）数値条件の下限値と上限値が同じ桁数の場合

数値の分割は行わない。

以上のように，数値条件を分割することにより，分割された各数値条件は特定の桁数の数値になるため，それぞれのオートマトンが簡単になる。

具体的に，

１２　≦　Ｋ　≦　３４５　　（１）

の数値条件を例に取り説明する。この例では，数値条件の下限値と上限値の桁数が１桁異なるので，（２）の分割方法に従い，次の２つに分割する。

１２　≦　Ｋ１　≦　９９　　（２）

１００　≦　Ｋ２　≦　３４５　　（３）

オートマトンは，図１のように極めて簡単になるため，オートマトン生成時間が短くなる。このオートマトンは，分割条件ごとにマイクロプロセッサを用いて並列に処理することによって，高速化が実現できる。

3．おわりに

数値条件分割方式を提案した。この方式の特徴は，以下の通りである。

（１）オートマトンが簡単になり，オートマトン生成時間を短くできる。

（２）分割した数値条件ごとに並列に照合処理することにより，高速化が実現できる。

【参考文献】

[1]L.A.Hollar,"Text Retrieval Computers," Computer,Vol.12,No.3,Mar.1979,pp.40-50.

[2]A.V.Aho and M.J.Corasick,"Efficient String Matching,"Communications of the ACM,Vol.18,No.6, June 1975,pp.333-340.

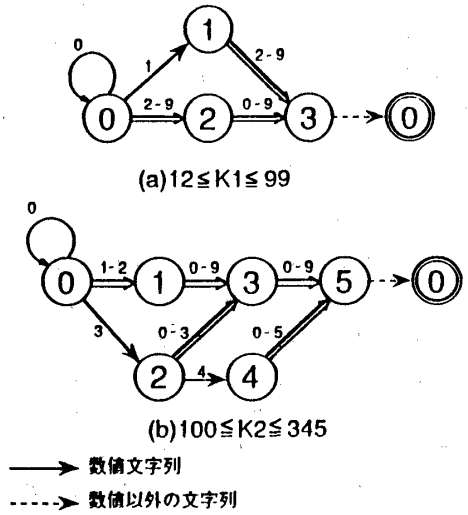

図１．数値条件分割方式の状態遷移図(12≦K2≦345)

Proposal of Range Search Engine for Text Search

Takanori Shimura,Hisamitsu Kawaguchi,Kanji Kato,Atsushi Hatakeyama,Mituru Akizawa
HITACHI,Ltd.