2N-6 自由語による全文検索のためのテキストサーチマシンＴＳＭ－Ｉ

加藤 寛次，藤澤 浩道，川口 久光，畠山 敦，大山 光男，兼岡 則幸
(株)日立製作所 中央研究所

１．はじめに

近年，ワードプロセッサやパーソナルコンピュータ，ワークステーションなどの普及拡大に伴い作成される文書情報も急速に増加してきており，近い将来膨大な量に達するものと予想されている。このため，大量の文書情報を一般のユーザが簡単に蓄積・検索できる文書情報検索装置に対する要求が高まりつつある。

こうした要求に応えるものとして，インデックス情報を用いない自由な言葉による検索を目的としたフルテキストサーチ(自由語全文検索)技術の研究を行なってきている。ここでは，そのためのテキストサーチマシンＴＳＭのアーキテクチャと試作したプロトタイプについて報告する。

２．フルテキストサーチの課題

自由語によるインデックスを用いないフルテキストサーチを実現するためには，以下に示す課題を解決する必要がある。

（１）ユーザが指定したキーワードとテキスト中に記述されたキーワードの記述・表記形態の違いによる検索漏れの防止。

（２）実用上許容し得る時間内で検索を実行できる高速なスキャン型検索処理の実現。

（３）小規模システムから大規模システムまで自由に構成でき，かつデータベースの増加にも容易に対応できるシステムアーキテクチャの実現。

３．フルテキストサーチ方式

上記の課題に対して，本テキストサーチマシンＴＳＭでは以下のようなフルテキストサーチ方式を採った。

（１）記述・表記違いの検索漏れの防止

ユーザが指定したキーワードの同義語と異表記語をシステム内部で自動的に生成して，これらをまとめてキーワードとして検索する同義語・異表記検索方式の採用。

（２）高速なスキャン型検索処理の実現

①テキストデータの高速な読出し

複数台の磁気ディスク装置を並列に並べ，同時にテキストデータを読み出す集合型磁気ディスク(マルチディスクユニットＭＤＵ)制御方式の採用。

②高速な多重文字列サーチ

同義語や異表記語を含めて複数のキーワードを，テキストのただ１回の走査で一括

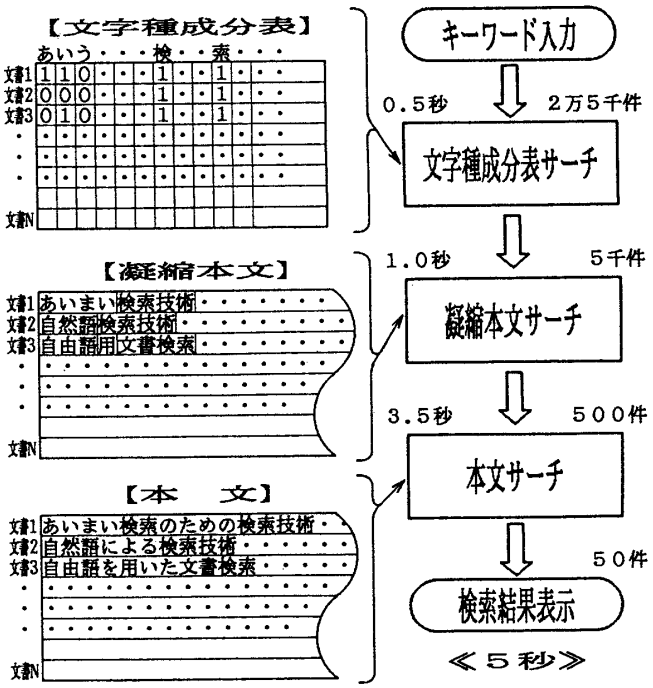

図１．サロゲート型検索加速方式

An Index-Free Full-Text Search Machine TSM-I
Kanji KATO, Hiromichi FUJISAWA, Hisamitsu KAWAGUCHI,
Atsushi HATAKEYAMA, Mitsuo OHYAMA, Noriyuki KANEOKA
HITACHI,Ltd.

してサーチできる有限オートマトン型検索プロセッサ(サーチエンジンＳＥ)の採用。

③フルテキストサーチの加速

あらかじめテキスト本文から自動的に作成，情報圧縮しておいた「文字種成分表」と「凝縮本文」を用いた階層的検索によって，検索速度を等価的に高めるサロゲート型検索加速方式の採用。（第１図参照）

④複合条件の高速判別処理

キーワード間の近傍，文脈及び論理条件に関する判別を一貫処理するパイプライン型複合条件判別方式の採用。

（３）システムの拡張性確保

それぞれ１台のサーチエンジンＳＥとマルチディスクユニットＭＤＵで検索ユニットＳＵを構成し，これを複数ユニットまとめて１台のサーチマシンＴＳＭとし，更にこのサーチマシンＴＳＭを複数台並列にＬＡＮに接続できるようにしたビルディングブロック型アーキテクチャの採用。

４．プロトタイプの試作

今回試作したプロトタイプは，１２台の小型磁気ディスクと１台のサーチエンジンＳＥで構成された１サーチユニットＳＵであり，これは２０５０／３２を介してＬＡＮ(CSMA/CD)へ接続される構造になっている。マルチディスクユニットＭＤＵの最大読出し速度は１０ＭＢ／ｓ，サーチエンジンＳＥの最大照合速度は２０ＭＢ／ｓである。（第２図参照）

このサーチユニットＳＵには平均２０ＭＢの容量の文書が２万５千件格納でき，約１千語の同義語・異表記語を含めて約５秒で検索できる。等価的には１００ＭＢ／ｓの検索速度が得られる。

これを用いれば，約２年半分の新聞記事が５秒で検索できることになる。

５．おわりに

サロゲート型検索加速方式を用いることによりスキャン型のフルテキストサーチを実用的な速度で実現できることを確認した。

今後は，実データに基づきフルテキストサーチの有効性を検証して行きたい。

参考文献

（１）畠山 他，自由語検索のための同義語・異表記展開方式，第３９回情処全大

（２）川口 他，自由語検索のための高速文字列検索方式，第３９回情処全大

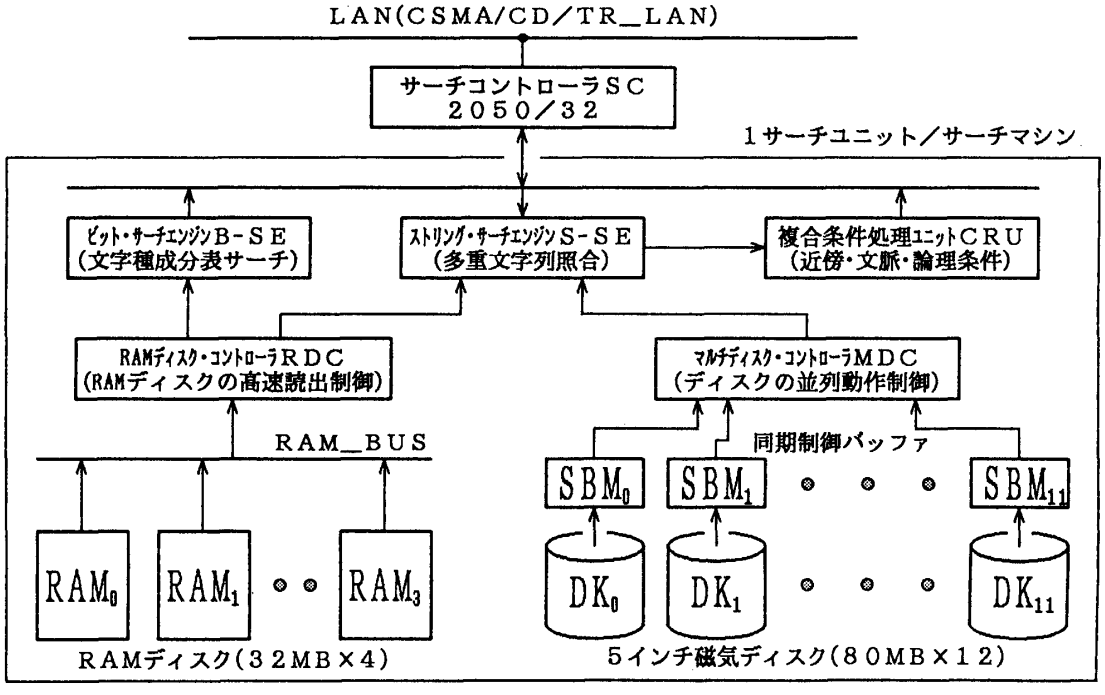

図２．試作テキストサーチマシンの構成