計算機アーキテクチャ　90－7
（１９９１．１０．７）

テキストDBマシンと検索プロセッサのアーキテクチャ

高橋恒介　　山崎哲矢、　　西塚博文　　本村真人
日本電気（株）、C&Cシステム研究所、研究開発技術本部,マイクロエレクトロニクス研究所

ここではマルチプロトコル文書データの検索を高速化するデータベース（DB）マシンアーキテクチャと入れ子構造で階層的なデータの検索プロセッサのアーキテクチャを提案する。文書データの多様化、マルチプロトコル化に対応して、シソーラス、文書データをOSIメッセージのプロトコルASN.1で記述し、シリコンデバイスを使う高速補助記憶（FAM）に格納する。FAMに直結されるOSIメッセージ検索プロセッサMSPと文字列検索プロセッサDISPによりテキストデータ部分の選択的検索を高速化する。実現性の確認のために開発されたMSPとDISPのLSIを使用するDBマシンのシステム構成と動作が検討される。

Architecture for Textual DB Machines and Search Processor LSIs

Kousuke Takahashi, Testuya Yamazaki, Hirofumi Nishizuka, and Masato Motomura

C&C System Research Laboratories, R&D Planning & Technical Service Division
and Microelectronics Research Laboratories of NEC Corp.

4-1-1, Miyazaki, Miyamae-ku, Kawasaki 216, Japan

This paper proposes textual DB machine architecture and nested data search processor architectures to accelerate the multi-protocol document retrieval. Document and thesaurus data are described by the rule ASN.1 for the OSI messages, stored in the fast auxilliary memory (FAM) using silicone file devices, and searched by a combination of an OSI message search processor MSP and a string search processor (DISP), which have been realized as LSI chips. The DB machine system using LSI chips is also discussed.

1、まえがき

テキスト検索の対象を文書一般に広げると、それは文字列以外に、文字でない部分（写真、グラフ、表、図形）を含むマルチメディア情報であるが、メディア毎でのデータの作成され方の不統一を考えると、文書はマルチプロトコル情報となる。そのような情報の検索をここで考える。

図面や表を含む文書をイメージデータとして記憶している図１の電子ファイリル検索システムでは、従来、データ検索が文書の書誌データを貯えるインデックスファイルやテキストファイルの検索で高速化されてきた。ソフトウエアによる高速化に限界があって、専用ハードウエアを使うＤＢマシンが考え出された。図１（ｂ）のようにインデックスファイルの検索を高速化するＤＢマシン、図１（ｃ）のようにテキストファイルの全文検索を高速化するＤＢマシン、図１（ｄ）のようにシグナチャファイルのような圧縮テキストを作ってその検索を高速化するＤＢマシンも開発された。

情報検索を問題解決に必要な知識獲得行動と捉えると、書誌データを含むインデックスファイルとテキストファイルとシソーラスは一体にして検索できるようにすべきである。その時の問題は、ファイルデータのプロトコル（日本語／英語／外国語、ビット列／コード列、構造／非構造、圧縮／暗号符号、図／表など）の違いへの対応にある。通信ネットワークを介してファイルデータが交換されるような情報社会が到来するとすれば、文書ＤＢ設計にもＯＳＩの階層アーキテクチャを念頭におくべきとの指摘がある[2,3]。

異質な文書データがメッセージ列に統合されることによって文書検索が広く役立つものと考えられる。そこで、これまでの一様なテキストデータの検索から、マルチプロトコルの文書データ検索への拡張のアプローチを提案する。

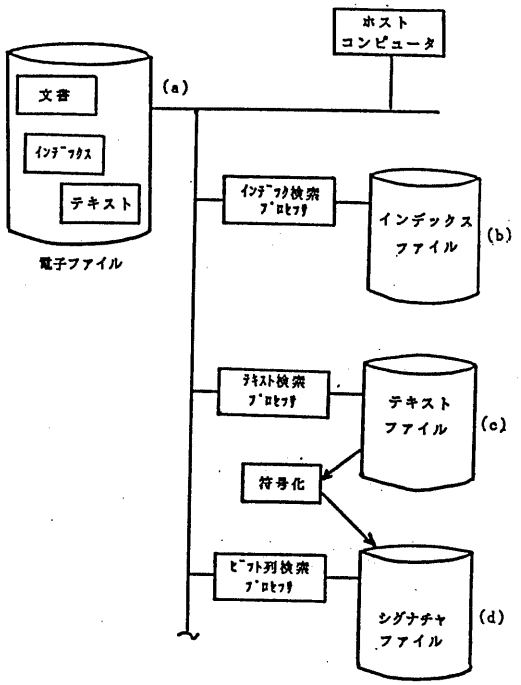

図１、電子ファイルの検索システム例

2、マシンアーキテクチャ

ここで提案するテキストＤＢマシンは、図２に示すように、高速補助記憶（ＦＡＭ）と２段構成の検索プロセッサの組み合せで、書誌データ、シソーラスを含めた文書データのテキス部分の高速検索を可能にする。図１（ｂ、ｃ、ｄ）に代るものとなる。構成を簡単にする代りに文書データは完全な非構造から僅かだけ構造化されたものになる。

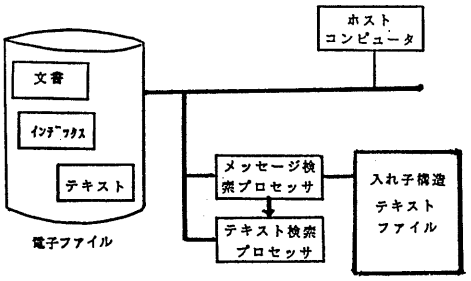

図2、テキストDBマシン

このようなＤＢマシンはコンピュータネットワークの端末側で使われ、電子ファイルはネットワーク・サーバ側に用意されるとする。ユーザは図2のＤＢマシンに格納されたマルチプロトコル・マルチメディア文書情報を利用し、問題解決に必要な知識の獲得を行なう。

このＤＢマシンの設計概念の概略は、

1、構造化されたインデックスファイルやシソーラスと非構造のテキストファイルを統合する事、
2、ファイルデータをＯＳＩプロトコル（ASN.1）で記述し、マルチメディア文書の出版・交換への対応の可能性を残す事、
3、ファイルデータを入れ子で階層構造のＯＳＩメッセージで合成し、メッセージ毎のプロトコルをメッセージ検索プロセッサで識別する事、
4、メッセージのフィールド別に検索条件を用意し、マルチフィールドの検索条件でファイルデータを一括検索する事。

ここに、フィールドは技術分野や文書の内容の分野、メディア、インデックスの属性、プロトコル別処理機能、オブジェクトのクラスなどである。

図3は、従来の一様なテキストデータの検索（ａ）と、上記設計概念により入れ子で階層構造を含むマルチプロトコルファイルデータの一括検索（ｂ）との違いを示す。

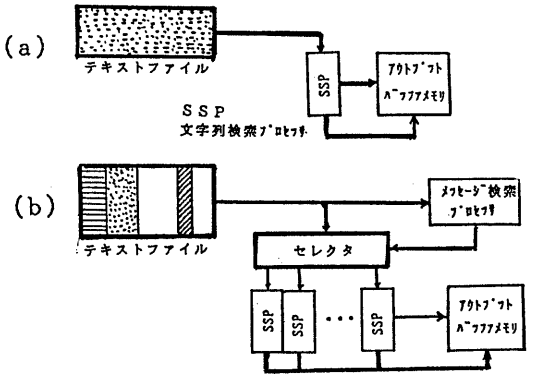

図3、テキスト検索の旧方式(a)と新方式(b)

図３（ｂ）は指定フィールド別に文字列検索プロセッサＳＳＰを用意し、文字列検索動作の前にフィールド識別を行ない、その結果で指定フィールドのテキストデータを該当するＳＳＰへ導くことを示す。これを行なうために、ファイルデータの内容（ＯＳＩメッセージ）を解読し、メッセージの先頭でそれのフィールドコードを検出するプロセッサを用意する。

図３（ｂ）を実現する時の主な技術課題は、

（イ）テキストデータの入れ子構造ＯＳＩメッセージへの変換ソフトと記憶管理ソフト、
（ロ）入れ子構造メッセージの選択的読み出しと、フィールドコードのリアルタイム出力を行なうメッセージ検索プロセッサ、
（ハ）フィールド別検索条件の設定可能でフィールド照合と文字列照合を並行して行なう文字列検索プロセッサの開発
（ニ）フィールド別シソーラス（用語辞書）の一般用と個人用の準備、である。

シソーラスは一般用、個人用で異なる知識情報の一種であり、情報検索は検索者の問題解決に必要な知識情報を獲得する１プロセスである。質問文を与えて、知識を個人用シソーラスに書き加えることである。

検索者は質問文から一般用シソーラスを使って検索条件を抽出し、それを検索プロセッサに登録して、辞書検索で意味や同義語や関連語を知り、文書のテキスト検索で使用例を知る。検索結果を検索文字列と対応づけて個人用シソーラスに組み込む。

個人別シソーラスは一般用シソーラスが階層構造化されていれば、同じような階層構造になる。文書検索回数を重ねる事を通して、個人シソーラスを充実させる事が知識獲得に相当し、新旧検索条件を個人用シソーラスをベースに比較照合し、全文走査無しで検索結果を得る確率を高めることが性能を改善する学習効果に対応する[7]。

なお、ＤＢマシンはイメージプロセッサの付加によってテキストデータ以外のデータも処理可能とする。ＭＳＰがテキストでないデータを区別し、専用メモリを介しイメージプロセッサにデータ振り分けるからである。この点はここでは議論しない。

3、ファイルデータのデータ構造

ここでは、異機種コンピュータ間での文書データ交換に使われる抽象構文規則ＡＳＮ．１を利用し、文書データをＯＳＩメッセージの集合として記述する。マルチメディア・マルチプロトコルの文書要素データをプロトコルが個別に指定されるメッセージで表し、階層的に関連する文書要素データを入れ子（階層）構造で連結する。結果はバイトシリアルなメッセージ列となる。

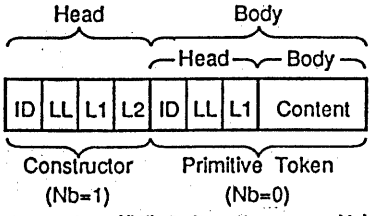

図４、入れ子構造ＯＳＩメッセージ例

ＯＳＩのメッセージは、図４のように文書要素データをボディとし、その前にボディへの処理機能を示す識別子（ＩＤ）と長さコード（ＬＬ、Ｌ）から成るヘッダを付けたものである。識別子の上位３ビット目のネストビットが'1'か'0'かによってボディ部に入れ子を持つか否かを示す[4]。ヘッダの付加は、情報通信ネットワークを使って文書交換を行なう時のフォーマットに準じており、かなり緩やかな構造化と言える。

ＯＳＩメッセージの合成は図５のように行なえる。（ａ）は文書データが一様なテキストデータだけであるときのメッセージであって、先頭に１つだけヘッダ■を持つことを示す。（ｂ）は書誌データと本文から成る文書データが２つのメッセージの先頭にコンストラクタと呼ばれるヘッダ□を持つことを示す。（ｃ）は書誌データに対応するメッセージが入れ子で属性別のデータを含むことを示す。（ｄ）は本文に対応するメッセージも入れ子で色々な文書要素を含むことを示す。文書要素が写真図面であってもよい。

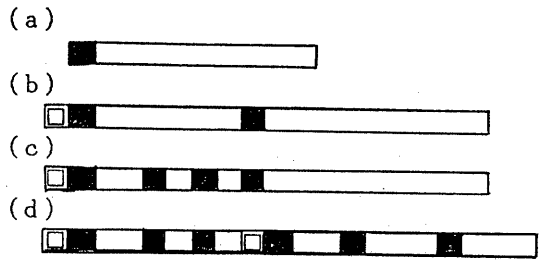

図５，文書データの入れ子による階層構造化

文書要素データを入れ子構造のメッセージ列に変換する手順は、文書要素データをカッコ（または）とスラッシュ／でクラス分けし、（の数でネストレベルを指示するか、文書要素の境界を改行で、階層を改行後の文字列印字開始位置で指示するか、である。長さコードは文字数の計算によって決る。

図５（ａ）のように、ヘッダが１つ付加されただけでも、ヘッダ部のＩＤコードの下位５ビットでボディ部のクラスを任意に分類できる。１ビットをイメージと文字コードの区別に使い、残り４ビットを情報の分野の識別に使うとしても、１６フィールドの文書要素をカバー出来る。

図５（ｃ、ｄ）のように、構造化されたＤＢのインデックスファイルを文書データと別に含む場合には、入れ子を２層にする。ここでは、１階層目で１６種のデータが記憶され、２階層目の中に１６フィールドに分類されたデータが含まれる。

入れ子の深さが増すと、メッセージ列は図６のようなシンタックストリーと対応づけて読まれる。このときの階層の高さをネストレベルＮＬと呼ぶ。各文書または文書集合の構造が複雑であっても、図６（ｂ）のように分類できれば、（ａ）のようなメッセージ列に変換できる。このことは、文書データだけでなく、書誌データ、シソーラス、検索結果などのデータ記述についても同じである。

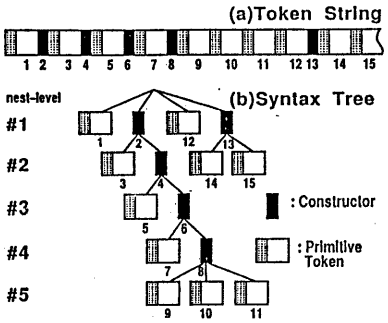

図６、メッセージ列と階層トリーとの対応付け

文書データや書誌データやシソーラスや検索結果をメッセージ列に変換したあとは、バイトシリアルなファイルデータとしてＦＡＭに記憶する。以下はこのようなメッセージ列の検索システムを考える。

４、ＯＳＩメッセージ検索プロセッサ

ＯＳＩメッセージ列の検索には図７に示すようなメッセージ検索プロセッサＭＳＰが必要となる。入力はメッセージ列であり、出力はメッセージ列をＦＡＭから呼び出すアドレス、メッセージ列のヘッダ部の解析結果（ＩＤ、ＮＬなど）である。

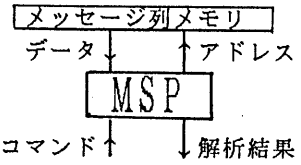

図７、ＭＳＰのＬＳＩチップのイメージ

メッセージのヘッダ部の解析は各メッセージの先頭がヘッダ部のＩＤコードであること（図４）をベースに、図８（ａ）の順序論理に従って行なわれる。１バイト目でＩＤとネストビットＮｂを、２バイト目で長さ形式ＬｂとＬコード数を、３バイト目以降でＬコードを読み取り、ボディ部長さＢＬや次のメッセージの開始位置（ＨＡ）を計算する。その間にＩＤ、ＮＬ、Ｎｂの照合を行ない、照合結果Ｚｂで次に読み出すアドレスを計算する。

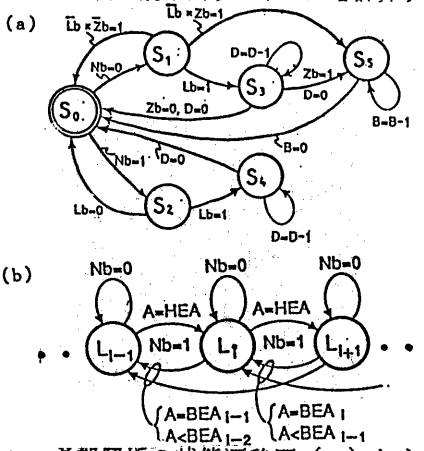

図８、ヘッダ部解析の状態遷移図（ａ）とネストレベル判別のための状態遷移図（ｂ）

ネストレベル判別には図８（ｂ）の状態遷移図が使われる。Ｎｂが１のＩＤコードを持つヘッダ（コンストラクタ）がきた時に入れ子のメッセージが続き、そのメッセージのＮＬが１つ高くなる。入れ子のメッセージを読み終わると、ＮＬは元の高さにもどる。入れ子メッセージの中に入れ子メッセージが続くとＮＬは１段づつ高くなるが、深い方のメッセージが読み終わり次第、ＮＬは下がる。何段もまとめて下がる場合もあるため並列処理が必要となる[5]。

メッセージ検索プロセッサＭＳＰは図８（ａ）の状態遷移図に従った順序論理回路、ミスマッチメッセージのボディ部を読み飛ばすアドレス計算回路、図８（ｂ）に従って次のメッセージのＮＬを決定するネストレベルモニタ回路、検索モード選択回路から成る。図９は72ピンＰＧＡに実装されるＬＳＩチップを示し、表１はＭＳＰのＬＳＩチップの諸特性を示す[5]。

検索モードや検索範囲などのコマンドは始めに設定される。ＭＳＰの検索動作モードは次のとおりである。

1,実時間検索モード：ヘッダ部を除いた全文検索
2,ヘッダ検索モード：ヘッダ部だけの検索
3,断続的検索モード：メッセージ単位で検索停止
4,スキップ検索モード：マッチメッセージの検索

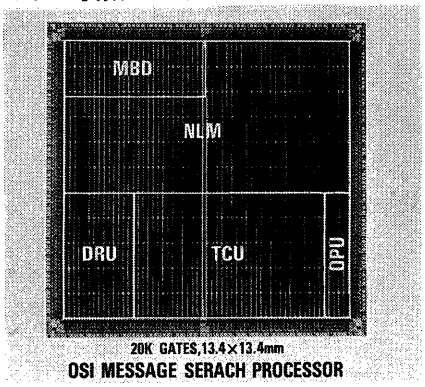

図９、メッセージ検索プロセッサＭＳＰ

表１、ＭＳＰのＬＳＩチップの諸特性

最大検索範囲	３２ビットアドレス空間
検索処理速度	２８０万オクテット／秒（５．６ＭＨｚ）
入力データ	ＯＳＩプロトコルＸ．４０９又はＡＳＮ．１で記述のもの
出力データ	アドレス／マッチドトークンＨＡ、ＢＡ、ＩＤ／ＮＬ、ＢＬ
使用電源単一	５Ｖ+10%、約７５ｍＡ（5MHz）
LSI ﾊﾟｯｹｰｼﾞ	７２ピンＰＧＡ、ＴＴＬ

５、高速補助メモリＦＡＭ

ＭＳＰが発生するアドレスに従ってメッセージ列のファイルデータを読み出せるＦＡＭとしては大容量・低コストで、低電力のＲＡＭが望ましい。シリコンファイルはＤＲＡＭをポータブルコンピュータ用に改良したものでありＤＲＡＭ並みの価格で消費電流が小さい。非動作時をオートリフレッシュモードにし、データ保持電流が３０ｕＡにできる。３００チップなら約１０ｍＡになる。この消費電流は１ＡＨの電池で１００時間のデータ保持を許す。停電や電源故障の期間のデータ保持には十分である。なお、３００チップは４ＭｂＲＡＭの場合、１Ｇビットに相当する。

シリコンファイルでは多数のチップを使っても同時に動作するチップが少ないので動作時の消費電流も小さい。１Ｇｂ当りで１Ａ以下となる。データ転送レイトは５～２０ＭＢ／ｓ以上であり、ページモードでは更に高い。フラッシュメモリに較べると書き換え動作が読み出しと同じ速度で実行されるメリットがある。磁気ディスクに較べると、ＭＢの価格が１桁以上高いが、性能価格比で見るなら、すでに,シリコンファイルが有利になる。そこで、以下ではＦＡＭにシリコンファイルメモリＳＦＭを採用することとする。

6、文字列検索プロセッサ

文字列検索方式としてＡＭ法、ＣＡ法、ＦＳＡ法、ＤＰ法とＰＳＬ法があるが、ＬＳＩ化にはＰＳＬ（ﾌﾟﾛｸﾞﾗﾏﾌﾞﾙ順序論理）法が良い。連想メモリＣＡＭに順序論理回路ＳＬＣを結合させたＰＳＬ法の文字列検索プロセッサＳＳＰをＬＳＩ化すると、可変長の文字列照合、アンカー／ノンアンカー文字列照合、固定長・可変長ドントケア文字列照合、ワイルドカード文字列照合、あいまい文字列照合の機能が可能になり、かつ、５１２文字の文字列登録と、６４以下の検索文字列の並列照合と、２０ＭＢ／ｓでの文字列検索も可能になる[6,7]。

しかし、このままの文字列検索プロセッサでは階層構造をもったファイルデータの検索に使えない。ＯＳＩメッセージ列のファイルデータから指定プロトコルのメッセージに含まれるテキストデータを選び出せば、ＳＳＰで検索が可能になる。

メッセージ検索プロセッサＭＳＰはヘッダ部を検出し、特定フィールド（プロトコル）のテキストデータだけを選択してくれる。問題は、図３（ｂ）に示したように、複数のフィールドにまたがる検索文字列により複数フィールドのテキストデータを１回のスキャンで検索させられないかという事である。

ｍ個のＳＳＰを１チップで実現するＬＳＩ設計に際しては、単にＬＳＩ規模を高めるだけでなく、ＳＳＰのＣＡＭ部のみをｍ個に増やしＳＬＣ部をそれらで共用させるＶＬＳＩアーキテクチャと、ｍ個の中の１つのＣＡＭの選択信号を、メッセージのフィールド（ＩＤ・ＮＬ）照合によって発生するマッチ回路の内蔵が望まれ、図１０に示すような辞書検索プロセッサＤＩＳＰが開発された[8,9]。文字記憶容量が１６倍に増え、１チップが１６個分のＳＳＰに相当する。それにも係わらず、検索速度がＳＳＰの場合以上と報告されている。

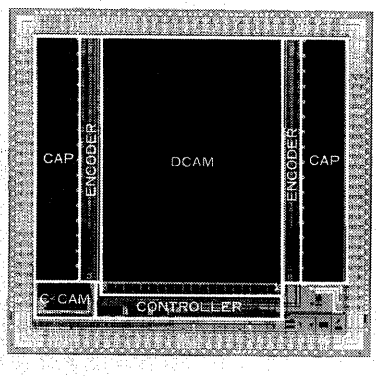

図１０、ＤＩＳＰのＬＳＩチップ写真[8,9]

ＤＩＳＰのアーキテクチャはＣＡＭ部をＳＲＡＭで実現し、ＳＲＡＭの大容量化に合わせ、同じチップ面積でＣＡＭ部の記憶容量を増大可能にし、今後の大容量ＦＡＭや大規模シソーラスへの対応を容易にする。なお、ＤＩＳＰは複雑な階層構造の用語辞書を検索するが、辞書検索専用ではない。

6、ＤＢマシンのシステム

したがって、マルチプロトコル文書データ中からテキストデータ部分を選び出してテキスト検索を可能にするＤＢマシンのシステムは、図１１に示すようになる。ＳＦＭを用いたＦＡＭとそれをアクセスするＭＳＰとＦＡＭから読み出されたデータを検索するＤＩＳＰと、検索結果の出力バッファメモリＯＢＭが主な構成要素である。

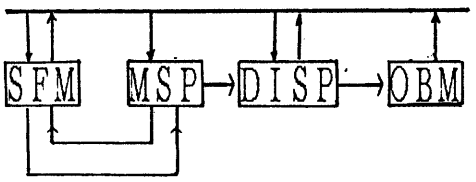

図１１、テキストＤＢマシンのシステム構成

実際にはＯＢＭの後に問い合わせ解読プロセッサＱＲＰが必要で、検索結果がＯＢＭに集められた後に、ＤＩＳＰへ検索条件式を入力し、それに対応したＯＢＭの出力をＱＲＰで処理する。

このシステムで行なわれる検索処理は、質問文と一般シソーラスとの照合、シソーラス間の用語文字列照合、シソーラスと文書データの照合、文書集合検索、引用文献欄の検索、個人シソーラスの検索などである。

検索速度の高くなる理由は以下のようになる。

（１）指定フィールドに含まれる各文書データの指定部分を検索する場合に、指定されない部分をスキップしながら、ＦＡＭの中のバイトシリアルなメッセージ列を検索すること、

（２）複数のフィールドの文書を１度で検索出来ること、

（３）問題解決に必要な用語を個人用シソーラスの検索で確認し、無駄な走査を省けること、

（４）ＦＡＭやＤＩＳＰやＭＳＰが高速に動作すること、

７、結論

マルチプロトコルの文書データをＯＳＩに準拠した抽象構文規則ＡＳＮ．１で記述することを考え、そのような文書データの検索を可能にするテキストＤＢマシンと検索プロセッサのアーキテクチャを提案し、検索プロセッサＭＳＰ、ＤＩＳＰのＬＳＩ開発結果を示した。ＦＡＭには低電力のシリコンファイルメモリが適当である事を示した。

ＤＢマシンは単純な構成で、多様な文書データ、シソーラスの高速検索の可能性を与える。今後の課題はＳＦＭ、ＭＳＰ、ＤＩＳＰを組み合せたシステム実験で具体的に情報検索を評価することである。

最後に、本研究の機会と御討議を頂いた山本昌弘所長並びに関係者各位に謝意を表す。

＜参考文献＞

1,C.W.Bachman;'A Personal Chronicle:Creating Better Information Systems with some Guiding Principles',IEEE Trans.Vol.KE-1(1),Mar.1989

2,加藤・藤澤、その他「全文検索用テキストサーチマシンの開発」、電子情報通信学会、技術研究報告DE89-38、Dec. 1989

3,I.A.Macleod;'A Query Language for Retrieving Information from Hierarchic Text Structure', The Computer Journal, Vol.34(3),p254, 1991

4，高橋、中川路「構文定義用言語ＡＳＮ．１の特質と処理系の現状と動向」情報処理 Vol.31(1), Jan. 1990

5，高橋、高橋、山崎「マルチメディア文書の記憶・検索とプロトコル解析ＬＳＩ」、電子情報通信学会、研究技報ＣＰＳＹ９０－９１、Jan.1991

6， K.Takahashi, H.Yamada and M.Hirata ; 'A String Search Processor LSI', JIP, J. of Information Processing, Vol.13(2), 1990

7，高橋、永井、山田.「テキストデータベースの学習型検索方式」電子情報通信学会、ＤＥ88-3、Ｍａｙ 1988

8,M.Motomura,etal;'A 1.2-Million transistor, 33MHz, 20-bit Dictionary Search Processor with a 160kb CAM,'Digest of Technical Paper, TAM5.4, IEEE ISSCC'90, Feb. 1990

9，本村、その他「１２０万トランジスタ辞書検索プロセッサ（ＤＩＳＰ）ーその構成と機能ー」、1990電子情報通信学会春期全国大会C-662、p 5-2