情報処理
最前線

安倍 史郎　fabe@soc.se.fujitsu.co.jp
松浦 正卓　GDD02665@nifty.ne.jp
富士通（株）システム本部

全文検索機能を適用したインターネット時代の個人用情報管理ソフト

今やインターネットによって世界中から容易に情報を集めることができるが，個人の保有する情報の管理手段はいまだ未熟な状態にある．文字列照合方式の全文検索エンジンを適用したソフトウェアツールが，この課題に対する有効な解決策となり得る．本方式の情報管理ソフトウェアは索引ファイルの作成が不要なため，情報の登録・更新・検索が簡便に行えるという長所を持っている．1つの情報を分類情報，テキスト情報，数値情報，日付情報，添付ファイル情報などの属性情報で構成することにより，備忘録，情報索引簿，論文管理，ホームページスクラップブックなど利用範囲の広い個人用情報管理ツールとなる．

■はじめに

近年の情報化の進展，時間軸の短縮に伴って，個人が日常接触する情報の量は飛躍的に増加している．それらの情報の大部分は本人にとって価値のない情報ではあるが，手元にとどめておけば後々活用できるかもしれない有益な情報も多くある．そのような有益な情報を無為に見過ごしてしまえば，その情報が必要となったときに再度その情報を探し出す作業には大きな苦労が伴う．

私たちは，今やインターネットによって世界中から容易に情報を集めることができるようになったが，私たち個人の保有する情報の管理となると，はなはだ未熟な状態にあり，いまだ有効な手段を持ち合わせていない．

この課題に対し，手元のパソコンで手軽に使用できる個人用の情報管理ソフトが有効と考えられる．そのような目的を持ったツールの姿について，実装事例を交えて述べる．

■個人用情報管理ソフトに求められる特性

個人が接触する情報源は，書籍／雑誌，書類，マスメディア，口伝え，本人作成の情報（最近，その多くはパソコン内に存在する），本人の頭の中など多様である．それら従来の情報源に加え，近年はインターネットにより莫大な情報に接触することができるようになった．個人用情報管理ソフトは，これらすべての情報源から得られた情報を何らかの手段によって管理できる必要があり，そのツールには以下のような特性が求められる．

簡便性：私たちは，再利用確率の高い情報は手間をかけても蓄積するが，そうでない情報を蓄積しようとは思わない．再利用確率がどの程度までの情報を蓄積するかは，蓄積作業および取り出し作業の手間の度合いに応じる．したがって，本目的のツールは操作の簡便性が決定的に重要な特性となる．

高速性：簡便性と同様な根拠で，情報の蓄積および取り出し速度も重要な特性である．

汎用性：私たちが手に入れる情報

の媒体は多種多様であり，情報の形態も文書，数表，画像，音声など多様である．また，情報の種類も，電話番号や家族の誕生日のような日常的なメモから，予定表，業務日誌，懸案事項一覧のような日常業務のメモ，さらには，論文，法律，診療記録などの業務に直接かかわる情報まで，多様に存在する．目的のツールは，このような管理対象情報の多様性をすべて吸収できる汎用的なものでなければならない．

携帯性：目的のツールは，常に座右に置いて使用できるものでなければならない．したがって，高性能サーバでしか動作しないようなものであってはならず，モバイルパソコンに簡単にインストールできるハンディなものでなければならない．

再現性：蓄積された情報を再利用するための検索にあたっては，必要とする情報を可能なかぎり漏れなく取り出すための仕組みが必要である．

適合性：再利用の目的に適合しない情報を排除するための仕組みが必要である．

透明性：長い期間座右に置いて使用された目的のツールに蓄積された情報は，本人にとっては貴重な財産である．ベンダの都合で蓄積された情報を利用できなくなることがあってはならない．目的のツールは，誰もが蓄積された情報を取り扱える代替ソフトを開発できる，透明性の高いものでなければならない．

廉価性：個人利用であるので廉価でなければならない．また，可能な限り，他のミドルウェアを必要としないものが望ましい．

■個人用情報管理ソフトの要求仕様

以上の特性を満足する個人用情報管理ソフトが備えるべき仕様について述べる．

1件の情報を構成する属性情報

目的のツールの開発にあたり，汎用性の項で述べた多様な情報をいかにして1元的に取り扱うか，という問題をまず解決する必要がある．この問題は，1件の情報を以下の属性情報によって構成することにより解決できる．

分類情報

ツールの使用者の意志で，情報を分類登録できる機能を有しておく必要がある．この機能は，情報を取り出す時の再現性，適合性の面で有効である．

サブジェクト

本属性情報は単に登録情報のタイトルを記録するための項目ではなく，分類情報とは異なった視点で登録情報をグループ化するための情報である．たとえば，同一の文献から複数の読後メモを登録する場合に，本項目に文献名を記録しておけば，情報取り出し時に，指定文献から引用したすべての読後メモを引き出すことができる．顧客名，研究テーマなども同様な目的で本属性情報として使用され得る．

テキスト情報

本属性情報には文章表現の情報を登録する．本属性情報として登録する情報は，日記文や論文のような管理対象となる情報本体とは限らない．原情報は別に存在し，その原情報を探し出すために必要となる情報，たとえば，原情報の名称，キーワード，説明文，原情報の所在場所などを本属性情報として登録する．

日付情報

本属性情報は日付属性の情報である．アクション項目の解決期限，予定表の予定日，書類の保管期限など，登録情報の取り出し時に日付属性情報が重要なキーとなる情報も多い．本属性情報はそのような用途で利用する．

数値情報

本属性情報は数値属性の情報である．長さ，重さ，年代，価格など，登録情報の取り出し時に重要なキーとなる数値属性情報を，本属性情報として登録する．

添付ファイル

パソコン内には，画像，文書，表，HTML，プログラム，ホームページのショートカットなどさまざまな形態の情報が存在する．それらの情報を，直接，情報管理ソフトに取り込む必要はない．それらの情報を格納しているファイルのパスを添付ファイルとして登録しておき，その情報の名称，キーワード，説明文などをテキスト情報に登録しておけば，間接的に，パソコンで取り扱えるあらゆる形態の情報の管理が可能となる．

以上の仕様を実装した画面の例を**図-1**（「スーパーメモ帳 アラヤ識」[1]の例．以下の画面例も同じ）に示す．

情報の取り出し

検索

蓄積された情報の中から必要な情報を取り出すために，情報を検索する手段を有する必要がある．検索機能として備えるべき最低限の条件として，前述のすべての属性情報（添付ファイルを除く）について検索条

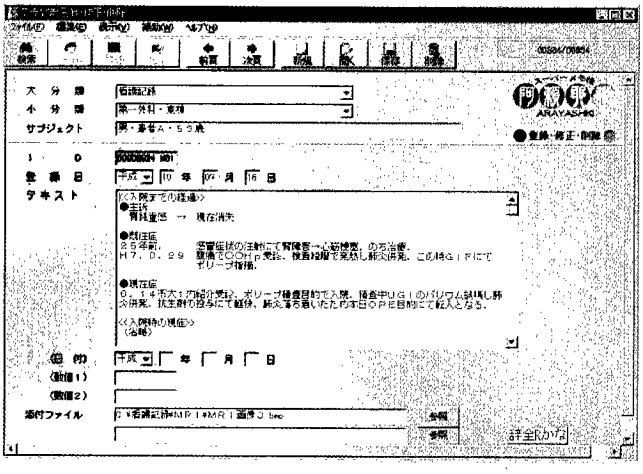

図-1　情報登録画面

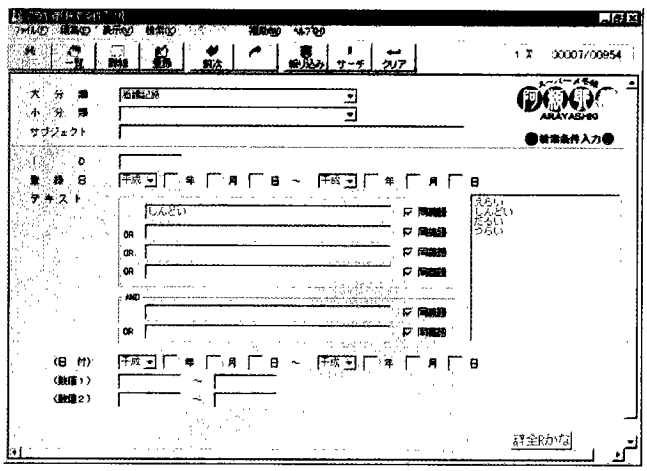

図-2　検索条件入力画面

件を指定できる必要がある．

日付情報および数値情報については，範囲検索機能は必須である．

登録されている情報で，アメリカ，米国，合衆国のように，同じ意味でも異なった表現がされている場合がある．したがって，再現性を高めるために同義語辞書の装備が望まれる．また，アルファベットの大文字，小文字，半角，全角の違いもソフトで自動的に吸収して，文字の形式の違いによる検索漏れを避ける必要がある．

以上の検索仕様を実装した検索画面例を図-2に示す．

検索結果の表示

検索条件を満足する情報の数が，十分に多い場合も考えられるので，まず，ヒット件数を表示して，適正な数になるまで絞り込みを行える機能を有する必要がある．

また，使用者が最終的に必要な情報を選択するための一覧表表示機能が必要である．一覧表は，全属性情報の一部を表示するなどの方法で，情報の内容を判別しやすくするための工夫が必要である．さらに，任意の属性情報をキーとして，一覧表表示情報の並べ替えができると，求める情報を探す手間もより少なくなる．

求める情報が絞られた後は，個別情報の表示機能が必要である．個別情報表示機能では，紙をめくるように前後の情報をワンタッチで表示できれば，操作の簡便さにつながる．

以上の仕様を実装した，一覧表画面例と個別情報画面例を図-3，図-4に示す．図-4は添付ファイルを開いた状態である．

情報の更新

検索表示した個別情報画面では，情報の変更，削除が簡便に行えなければならない．もし，それができなかったり，操作が複雑であった場合は，個人用情報管理ソフトとしては，著しく用途が制限され，冒頭で述べた目的を達成できないであろう．

情報の登録

情報登録機能として基本的に必要となる機能は，すべての属性情報を画面から単純に入力できる機能である．もちろん，コピー＆ペーストなどのOSの持っている基本機能は使えるものとする．しかし，それだけでは，簡便さの面で，個人用情報管理ソフトとしてのニーズに耐えきれないであろう．

そのニーズには，本稿で述べる個人用情報管理ソフトの周辺に，自動的に情報を取り込むためのツールを開発することにより応えることができる．たとえば，パソコン内のファイルから，テキスト部分をテキスト情報に，原ファイルのパスを添付ファイル情報に自動登録するツールが考えられる．WebサイトのHTMLを自動的に取り込み，自分のパソコン内にそのWebサイトのポータルを作るようなツールも考えられる．

利用者がこのようなツールを独自に作成することができるようにするには，透明性の項でも述べたように情報登録ファイル構造の透明性が確保されるだけで足りる．

■個人用情報管理ソフトの実装技術

以上の要求仕様を実装するために必要となる技術3点について触れる．

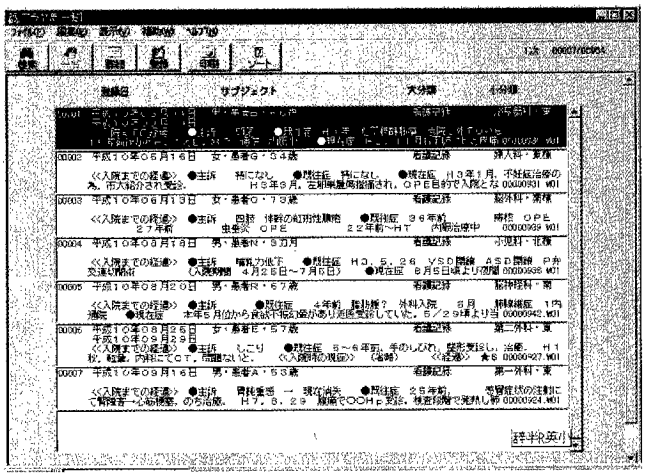

図-3　一覧画面

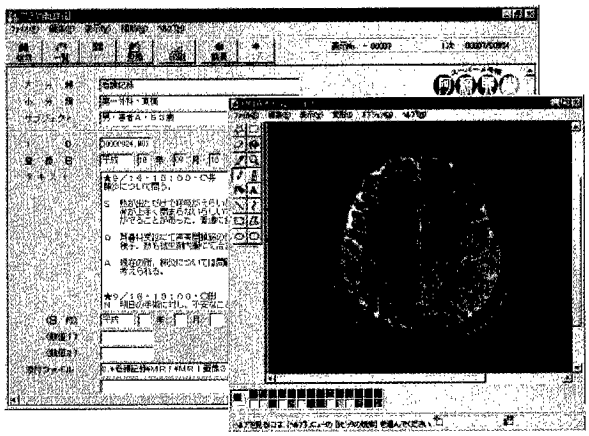

図-4　個別情報画面

検索手法の選択

目的のツールを開発するには，まず，蓄積された情報を検索するための検索手法を選択する必要がある．一般的な情報検索システムを評価しながら，最適な検索手法を選択する．

一般的な情報検索システムの評価

情報検索システムには，情報の主題概念をキーワードとして設定して，そのキーワードと原情報の在りかの対応をとるキーワード方式と，検索対象文書本文に含まれている全文字列をキーワードとする全文検索方式の2つの方式がある．

キーワード方式は，キーワードの設定に多大な手間がかかり，さらに，キーワードの設定作業は作業者の主観に負うところが大きく，検索者のニーズに合致するとは限らないという問題を含んでいる．一方，全文検索方式は，主題概念がキーワードとして漏れる可能性があるものの，検索者の工夫次第で漏れのない検索ができるという利点がある．また，本方式は，索引ファイルを自動的に作成することができるという長所があるため，現在は，検索システムの主流となっている．目的のツールの検索方式としては，情報登録操作の簡便性から判断して，全文検索方式を妥当とするのは異論のないことだと考える．

全文検索方式には，Nグラム方式や形態素解析方式など多数の方式が考案されているが，いずれも，検索速度を上げるために何らかの索引ファイルを作成する必要がある[2]．これらの検索方式には，索引ファイルが原文の数倍から十数倍となる，原文の登録／更新時の索引ファイル作成処理が運用面で大きな負担となる，索引ファイルを利用して検索するにはRDB相当のミドルウェアが必要となる，という問題がある．そのような問題から，これらの検索方式は，簡便性，携帯性の面で目的のツールには不向きである．

文字列照合方式の利点

そこで，注目すべきは，テキストエディタなどで昔ながらに採用されている文字列照合方式または一方向逐次処理方式と呼ばれる，対象文章を先頭より後方向へ逐次読みながら検索条件に合致する文字列を探す方式である[3),4)]．本方式には以下のような利点がある．

- 索引ファイルが不要であるのでファイル領域に冗長度がほとんどない．
- キーワードの抽出作業が不要のため情報の登録・更新・削除が自由である．
- RDBなどのミドルウェアを必要としないので低コストかつコンパクトに実装できる．

本方式にはこのような大きな利点があるにもかかわらず，検索時間の点で実用的でない，という理由で顧みられていない[5]．しかし，全文検索方式の研究開発にしのぎを削っている間に，今や，文字列照合方式でも利用分野によっては十分な性能を確保できるまでにハードウェア性能は向上している．

九州大学の有川教授らが開発した一方向逐次処理方式の検索エンジン「SIGMA」☆1を使用して15MBのテキ

☆1 SIGMAは，検索条件をオートマトン化することにより，検索対象のテキストを1回操作するだけで複数個の検索条件のパターンのマッチング処理を同時に行い，検索条件の数に比例しない高速検索を実現している検索エンジンである[6]．

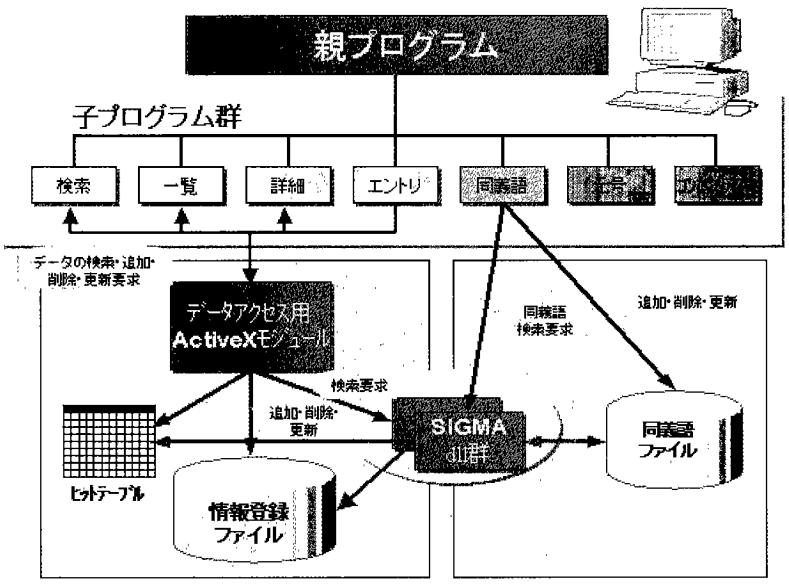

図-5　ソフトウェア構成図

ストを対象として実測したところ，検索時間は約1秒（Pentium IIプロセッサ 300MHzでの実測）であった．15MBのテキスト量は，1日400字で50年間の日記に相当する．また，1頁2,500字300頁の法律書10冊に相当する．検索時間は100MB程度まではほぼ線型に増加しており，ハード性能の向上，検索アルゴリズムの改善を考慮すれば，100MBを超えるテキスト量まで実用の範囲内に置くことができる．

この性能は，個人用情報管理ソフトとしてのニーズを十分満足していると考える．

ファイル構造

目的のツールのファイルは，透明性の配慮からプレーンなテキストファイルが望ましい．しかし，登録情報に要求仕様で述べた属性情報を持たせるには，テキストファイルを何らかの工夫で構造化する必要がある．前述のSIGMAの検索対象とするテキストは，1件の情報をレコード区切り語で区切り，さらに，その区切られた情報を項目区切り語で区切る構造となっている．

目的のツールのファイル構造は，この方式をそのまま踏襲することができる．それぞれの属性情報に対応する項目区切り語を設定し，情報登録時にはレコード区切り語，項目区切り語とともに情報を登録しておく．それにより，検索エンジンは，文字列照合検索の過程において，レコード区切り語，項目区切り語を認識した後，当該項目区切り語に続く情報がテキスト，数値，日付，添付ファイルのいずれの属性を持っているかを識別して，検索条件で指定された論理演算を行うことができる．

ソフトウェア構成

以上述べてきた仕様を実装した個人用情報管理ソフト「スーパーメモ帳 アラヤ識」のソフトウェア構成図を**図-5**に示す．本ソフトは親プログラムと7つの小プログラム群で構成されている．情報登録ファイルと同義語ファイルの検索はSIGMAを使用している．情報登録ファイルの操作機能は，ActiveXで構築し，部品化している．SIGMAの検索結果は，全レコードに対応したヒットテーブルを介して知ることができる．

■どのような使い方があるか

以上に述べた仕様で個人用情報管理ソフトができあがる．では，本ツールでどのようなことができるのか，いくつか具体例を挙げてみよう．登録情報の分類に応じて，各属性情報として登録する情報の種類を利用者が自由に定義できることから，本ツールを以下のように多面的に利用することができる．

情報保管庫

日記，読後メモ，アイディアメモ，技術メモ，予定表(日付項目を利用)，電子メール保管など，原情報を直接的に本ツールで記録管理しておく使い方である．

外部媒体情報の索引簿

書籍，雑誌記事，書類など原情報が紙などの外部媒体で存在する情報を管理したい場合，個人用情報管理ソフトは，それらの情報を迅速に探し出すための手段を提供すればよい．

たとえば，手元の書類を管理する場合，手元の書類に書類番号を付与して順に保管しておき，本ツールのテキスト情報には書類の名称・キーワードなど，日付情報には保管期限，数値情報には書類番号を記録する．このわずかな手間により，テキスト欄のフリーキーワード検索や保管期限での検索を行うことにより，書類の探し出しや廃棄の作業がきわめて容易になる．

書類番号を使用しなくて，小分類あるいはサブジェクト欄に保管場所を記録しておいてもよい．この方法は，図書管理，所蔵品管理など，原情報がコンピュータ外に存在する場合のすべてのケースに適用できる．

ホームページのスクラップ

探し当てたインターネット上の情報を手元のパソコンに保管しておきたいというニーズは大きい．しかし，一般的なパソコンの機能では，その作業はかなりの手間であり，保存した情報を後で探し出す手間も大きい．そのような情報を本ツールのファイルに取り込んでおけば，検索機能で容易に取り出すことができ，取り出した情報の添付ファイル欄をクリックすれば，原情報のホームページに接続することができる．

ファイル索引簿

多用されているパソコン内には数多くのファイルが保管されているが，それらのファイルを取り出す操作は，状況によっては手間のかかる作業となる．

本ツールの添付ファイル欄にファイルのパスを記録し，テキスト欄にファイルの名称，内容の概要，キーワードなどを記録しておけば，テキスト欄をフリーキーワードで検索した後，添付ファイル欄をクリックすれば，容易にファイルを取り出すことができる．

個人用情報ポータル

これらの用途以外にも業務での本格的な利用も考えられる．以上の使い方を整理した図表現を図-6に示す．本ツールは，原情報の保管庫として利用できるだけでなく，原情報の索引簿として効果的に利用できる．さらに，本ツールの分類機能と添付ファイル機能を活用することにより，本ツールは，個人の接触する情報の多様性に対応でき，個人の保有しているすべての情報の管理に利用できる．同じ操作で，全登録情報の横断的検索と全種類の情報の個別検索ができるため，操作は軽快である．本ツールを座右に置いて徹底して使えば，本ツールは，時間の経過とともに本人の保有する情報のポータルとなる．

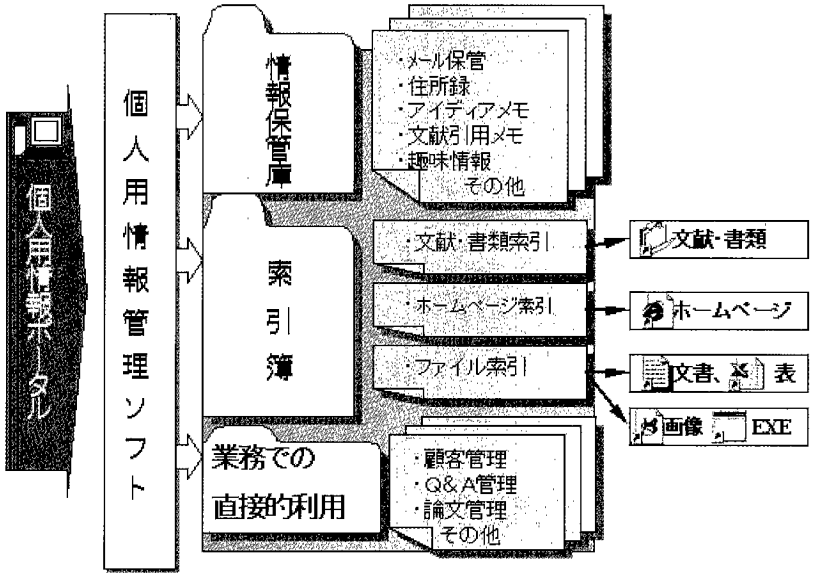

図-6　個人用情報ポータルとしての概念図

■将来の可能性

インターネット時代の到来により溢れる情報を個人的に管理する手段として，1つのツールを開発して，筆者なりに使い込み，その有効性について確認できた．

本ツールの情報登録ファイル構造の透明性により，第三者の，同ファイルへの情報登録ツールの開発，または異なった視点からの検索ツールの開発が可能である．また，辞書や法令などの電子出版物のコンテンツを同データ構造に合わせて出版すれば，出版社側で検索ソフトを用意する必要はなくなり，コンテンツの流通性が高まる．

筆者の実装したソフトには検索結果のCSV形式でのファイル掃き出し機能を持たせ，情報の再利用に配慮しているが，さらに，HTML形式やXML形式での掃き出し機能を持たせることにより，情報の再利用性がより高まると考える．

本稿で述べたような個人用情報管理ソフトが，情報化社会のリテラシーとなることを期待する．

参考文献

1) 富士通ミドルウェア（株），「スーパーメモ帳アラヤ識」，http://www.fmw.co.jp/
2) 全文検索システム協議会編，第3年度報告書，http://www.asahi-net.or.jp/~zc7t-urb/
3) ユン・ジヒ，高木利久，牛島和夫：5種類のパターン・マッチング手法をC言語の関数で実現する（第1回），日経バイト，pp.175-191（Aug. 1987）．
4) ユン・ジヒ，高木利久，牛島和夫：5種類のパターン・マッチング手法をC言語の関数で実現する（第2回），日経バイト，pp.233-243（Sep. 1987）．
5) 馬場　肇：日本語全文検索システムの構築と活用，ソフトバンク（1998）．
6) 有川節夫：テキストデータベース管理システムSIGMA第2版について，九州大学大型計算機センター広報，Vol.20, No.6（1987）．

（平成12年9月26日受付）